ISCTE IUL
University Institute of Lisbon

Department of Information Science and Technology

Efficient Solutions for Light Field Coding

Caroline Conti

Thesis specially presented for the fulfillment of the degree of

Doctor in Information Science and Technology

Supervisor:

Dr. Luís Eduardo de Pinho Ducla Soares, Assistant Professor

ISCTE - IUL

Co-supervisor:

Dr. Paulo Jorge Lourenço Nunes, Assistant Professor

ISCTE – IUL

December, 2016

ISCTE IUL
University Institute of Lisbon

Department of Information Science and Technology

Efficient Solutions for Light Field Coding

Caroline Conti

Thesis specially presented for the fulfillment of the degree of

Doctor in Information Science and Technology

Jury:

Dr. Ricardo Parreira de Azambuja Fonseca, Full Professor, ISCTE – IUL (President)

Dr. Frédéric Dufaux, Research Director, National Center for Scientific Research, France

Dr. Peter Schelkens, Full Professor, Vrije Universiteit Brussel, Belgium

Dr. Pedro António Amado Assunção, Coordinator Professor, Polytechnic Institute of Leiria

Dr. Luís Eduardo de Pinho Ducla Soares, Assistant Professor, ISCTE - IUL

December, 2016

Abstract

This Thesis aims at studying and developing efficient solutions for light field coding. In this regard, this Thesis proposes a light field coding solution based on the High Efficiency Video Coding (HEVC) standard and using a non-local spatial prediction scheme, named self-similarity compensated prediction. This solution is able to exploit the inherent correlations of this new type of content to achieve high rate-distortion performance, without requiring any explicit knowledge of the particular optical acquisition setup used when acquiring the content.

In addition to this, aiming at allowing 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 proposed in this Thesis. The proposed display scalable solution makes use of an efficient inter-layer prediction scheme that when combined with the self-similarity compensated prediction is able to achieve, in most of the cases, better rate-distortion performance than the non-scalable HEVC solution.

Finally, to support the richer and flexible interaction functionalities that arise in light field imaging applications, this Thesis also proposes a novel scalability concept, named Field of View (FOV) scalability, as well as a FOV scalable coding solution. The FOV scalability supports progressively richer interaction functionalities in each higher layer by hierarchically organizing the light field angular information. Moreover, two novel inter-layer coding solutions are also proposed so as to achieve high rate-distortion performance in enhancement layer coding.

Keywords: Light field, Holoscopic, Plenoptic, Integral imaging, Light field coding, HEVC, Coding efficiency, Display scalability, Field of view scalability

Resumo

Esta Tese visa estudar e desenvolver soluções eficientes para a codificação do campo de luz. Especificamente, esta Tese propõe uma solução de codificação do campo de luz que se baseia na norma HEVC (*High Efficiency Video Coding*) e que utiliza um método de predição espacial não-local aqui designada como predição compensada da autossemelhança. A solução proposta é capaz de explorar as correlações inerentes ao conteúdo de campo de luz e, desta forma, obter um alto desempenho em termos de débito-distorção sem que seja necessário conhecer os pormenores do sistema ótico utilizado na aquisição do conteúdo.

Adicionalmente, com o intuito de possibilitar uma difusão mais rápida de serviços e aplicações de campo de luz no mercado de consumo, esta Tese também propõe uma solução escalável de codificação do campo de luz de forma a garantir a compatibilidade com os dispositivos de visualização convencionais (por exemplo, 2D, 3D estereoscópico e 3D multi-vista). A solução escalável proposta assenta na utilização de um método eficiente de predição entre camadas que quando combinado com a predição compensada da autossemelhança consegue atingir, na maioria dos casos, um melhor desempenho em termos de débito-distorção que a solução não-escalável do HEVC.

Finalmente, para garantir o suporte das poderosas funcionalidades de interação que surgem em aplicações de campo de luz, esta Tese propõe um novo conceito de escalabilidade chamado escalabilidade de campo de visão, assim como uma solução de codificação de campo de luz que suporta este novo tipo de escalabilidade. A escalabilidade de campo de visão possibilita organizar hierarquicamente a informação angular do campo de luz de forma a suportar, progressivamente, funcionalidades de interação mais ricas em cada camada superior. Para alcançar um alto desempenho em termos de débito-distorção na codificação das camadas superiores, dois novos métodos de predição entre camadas são também propostos.

Palavra-Chave: Campo de Luz, Holoscópico, Plenótico, Imagem integral, Codificação do campo de luz, HEVC, Eficiência de codificação, Escalabilidade de visualização, Escalabilidade de campo de visão

Acknowledgements

First and foremost, I would like to thank my PhD supervisors Prof. Luís Ducla Soares and Prof. Paulo Nunes. Words cannot express how grateful I am for their outstanding guidance through this journey and for allowing me to grow as a research scientist. Their deep knowledge, integrity, patience and friendship make me sure that I could not have had better supervisors. I also thank Prof. Luís for making the reviews more exciting with his insightful and instructive comments flavored with his great sense of humor, and Prof. Paulo for his valuable advices mainly in tough times during this PhD pursuit.

I am grateful to all my current and former colleagues from the Multimedia Signal Processing Group of *Instituto de Telecomunicações*, who have contributed to a friendly and inspiring environment since I have arrived in Lisbon. A special thanks to Prof. Paulo Correia e Dr. Matteo Naccari for their willingness to help at the beginning of this journey, and Prof. Fernando Pereira for the interesting and fruitful discussions on light fields. I would also like to express my appreciation to Tereza Traquinas, Ana Rita Rodrigues and Sara Correia from *Instituto de Telecomunicações*, and Fátima Estevens from ISCTE-University Institute of Lisbon for their friendship and their valuable help in the administrative concerns. In addition, I would like to thank Prof. Tomás Brandão and Prof. Juan Acebrón for their support when I had to reconcile the teaching assistance schedule with the work on my Thesis.

My grateful thanks also to the participants of the 3D VIVANT project and the members of the 3D-ConTourNet COST Action for giving me the opportunity to learn and collaborate with experts from all over the world. I am also grateful to *Fundação para a Ciência e a Tecnologia* and *Instituto de Telecomunicações* for funding this PhD (under the SFRH/BD/79480/2011 grant and UID/EEA/50008/2013 project).

I would like to thank my friends who have always encouraged me. Special thanks to Roselaine, Marielle and Marcela whose friendship has proven to survive the time and the distance.

Last but not the least, this is a great chance to express my deepest gratitude to my family who has been my pillar of strength and to whom I dedicate this Thesis. Particularly, to my parents and sisters, for their dedication and love, and for supporting me in all my pursuits. To my aunties, for their encouragement and for helping funding our studies in the most difficult times. To Duarte for his patience and affection, and for his emotional support during the final stages of this PhD. To my grandmother Ramira and my aunt Maria, who are no longer with us, for their inspirational strength and perseverance in life.

Contents

List of Figures

List of Tables

List of Acronyms

1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
3DTV	Three Dimensional Television
4D	Four Dimensional
6DoF	Six Degrees of Freedom
7D	Seven Dimensional
AMVP	Advanced Motion Vector Prediction
AR	Augmented Reality
AVC	Advanced Video Coding
BD	Bjøntegaard Delta Metrics
Bi-SS	Bi-predicted Self-Similarity
bpp	Bits Per Pixel
BR	Bitrate
CABAC	Context-based Arithmetic Binary Coding
CB	Coding Block
CT	Computed Tomography
CTB	Coding Tree Block
CTU	Coding Tree Unit
CU	Coding Unit
DCT	Discrete Cosine Transform
DIBR	Depth Image Based Rendering
DPCM	Differential Pulse Coding Modulation
DS-LFC	Display Scalable Light Field Coding

DSCQS	Double Stimulus Continuous Quality Scale
DST	Discrete Sine Transform
DWT	Discrete Wavelet Transform
EBCOT	Embedded Block Coding with Optimal Truncation
EPI	Epipolar Plane Image
ES	Evolutionary Strategy
FN	Free Navigation
FOV	Field Of View
FOV-LFC	Field Of View Scalable Light Field Coding
FTV	Free-viewpoint Television
GPR	Gaussian Process Regression
GUI	Graphical User Interface
HD	High Definition
HDR	High Dynamic Range
HEVC	High Efficient Video Coding
HDM	Head Mounted Displays
HVS	Human Vision System
IL	Inter-Layer
ILR	Inter-Layer Reference
IntraBC	Intra Block Copy
ITU	International Telecommunications Union
ITU-R	Radiocommunication Sector of ITU
JCT-3V	Joint Collaborative Team on 3D Video Coding Extension Development
JCT-VC	Joint Collaborative Team on Video Coding
JP3D	JPEG 2000 Part 10 Standard
JPEG	Joint Photographic Experts Group
KLT	Karhunen-Loève Transform
k **-NN**	*k*-Nearest Neighbor
LBG	Linde-Buzo-Gray

LF	Light Field
LFC	Light Field Coding
LIDAR	Light Detection And Ranging
LLE	Locally Linear Embedding
MI	Micro-Image
MIVP	Micro-Image-Based Vector Predictors
MLA	Microlens Array
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MRI	Magnetic Resonance Images
MSE	Mean Squared Error
MV	Multiview
MVC	Multiview Video Coding
NAL	Network Abstraction Layer
NCC	Normalized Cross-Correlation
NHK	Nippon Hōsō Kyōkai (Japan Broadcast Corporation)
NMF	Non-negative Matrix Factorization
PB	Prediction Block
PCA	Principal Components Analysis
POC	Picture Order Count
PSNR	Peak Signal to Noise Ratio
PU	Prediction Unit
PVS	Pseudo Video Sequence
QP	Quantization Parameter
RD	Rate Distortion
RDO	Rate Distortion Optimization
RExt	HEVC Format Range Extension
RGB	Red, Green, and Blue
RPS	Reference Picture Set

SAD	Sum of Absolute Differences
SAO	Sample Adaptive Offset
SCC	HEVC Screen Content Coding
SMV	Super Multiview
SPIHT	Set Partitioning In Hierarchical Trees
SQ	Scalar Quantization
SS	Self-Similarity
SSD	Sum of Square Differences
SSIM	Structural Similarity Index
TB	Transform Block
TM	Template Matching
TU	Transform Unit
UHD	Ultra High Definition
Uni-SS	Uni-predicted Self-Similarity
VCEG	Video Coding Experts Group
VFX	Visual Effects
VI	Viewpoint Image
VQ	Vector Quantization
VR	Virtual Reality
WPP	Wavefront Parallel Processing
Y’CbCr	Luma, Blue-Difference Chroma, and Red-Difference Chroma

Chapter 1

Introduction

Over the last ten years, the world has witnessed a staggering progress in terms of multimedia devices, networks, and services. Each year, several new and more advanced devices (e.g., higher quality sensors and displays, and more computational power) are introduced into the consumer market. The combination of these advanced devices with the availability of higher bandwidth networks is leading to a widespread adoption of advanced multimedia applications and services.

Moreover, the evolution of mobile devices – such as smartphones and tablets – has completely changed the way people consume content, and there has been an ever-growing demand for visual mobile content. As reported by Cisco [1], more than a half of the mobile traffic is already video (see Figure 1.1), and it is expected to be three-quarters in the next four years. A relevant portion of this growth is due to the recent "social media revolution". Nearly any person can now be a content creator, capturing or creating his/her own images and videos, and instantaneously share or stream them on a social media platform.

The reason for this visual mobile explosion is also tied to the constant need for replicating the way Humans experiences the world and communicate. In this sense, there has been a continuous pressure for providing more immersive visual user experiences. Nowadays, advances in video acquisition and displays technologies allow supporting Ultra High Definition (UHD) spatial resolution, higher temporal resolutions and High Dynamic Range

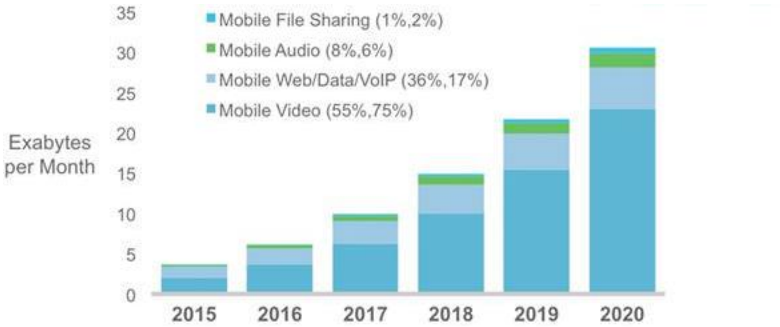


Figure 1.1 Cisco mobile data traffic projection for 2015-2020. In parentheses, it is shown that most of the data traffic is currently video (55 %), and it is expected that it will grow to 75 % by 2020. From: Cisco Visual Network Index Mobile, 2016 [1]

(HDR) contents. In this context, the most recent H.265 High Efficient Video Coding (HEVC) [2] standard has been playing a fundamental role in enabling efficient delivery of these new types of content to end-users over the infrastructure for broadcast television, mobile and internet services.

Following this continuous need for more realistic and immersive user experiences, the quest nowadays is: What is ‘the next big thing’? Although Three Dimensional (3D) stereo visual experiences had, just a few years ago, its moment of greater popularity in the Three Dimensional Television (3DTV) and 3D cinema markets, the public acceptance was far from the expected by the industry, and the percentage of display devices that support this capability has been drastically dropping in the consumer market these days [3, 4]. On the other hand, the research on alternative immersive acquisition and display systems has been rapidly progressing, as discussed in the following.

1.1 Richer Imaging Technologies and Prospective Applications

Richer imaging technologies have been rapidly maturing due to the recent developments in optics and sensor manufacturing [5]. In this context, the novel sensors, cameras and displays that have been emerging allow having richer forms of visual data with powerful new capabilities. Moreover, there has been a significant growing attention to these technologies in both research and industry areas, especially since the Oculus [6] company acquisition by Facebook [7].

In this context, the following three prospective imaging technologies can be highlighted along with their application domains:

- **Light Field** – Light Field (LF) imaging technology allows recording not only the spatial light intensity information of a scene but also angular viewing direction information. LF content can be captured by using an array of multiple conventional cameras or by using a single tier camera equipped with a Microlens Array (MLA) – also known as integral [8, 9], holoscopic [5], and plenoptic [10] imaging. Moreover, it can also be computationally generated using an LF camera model. LF imaging has become a prospective and practical approach, being applicable in many different areas, such as richer photography capturing [11–13], video production and Visual Effects (VFX) [12], 3DTV [5, 14–18], Free Navigation (FN) [19, 20], biometric recognition [21], medical imaging [9, 22], and Virtual Reality (VR) and Augmented Reality (AR) applications [23, 24]. Examples of LF applications for cinema production and FN can be seen, respectively, in Figures 1.2 and 1.3. It is worth already emphasizing here that this Thesis is focused on LF imaging applications that make use of the microlens-based acquisition setup.
- **Point Cloud** – Point cloud imaging technology allows recording a scene as a set of points in a 3D space associated with some additional attribute data, such as light color

Figure 1.2 Lytro Cinema [12] *(microlens-based) LF camera (left), and software demo (right) for VFX*

and direction information, and material reflectance properties. Point clouds can be captured by using 3D scanners and/or Light Detection And Ranging (LIDAR) cameras. Moreover, it can also be computationally generated from LF content or from 3D models. Point clouds have been specially applied to geographic information systems [25], autonomous navigation based on large scale 3D maps [26], cultural heritage archiving [27], as well as VR and AR applications (e.g., immersive telepresence [28], and social media [29, 30]). An example of an immersive social network application that might benefit from point clouds is depicted in Figure 1.4.

- **Holography** – In this case, a wave-based light propagation model is used in which the content, known as a hologram, records the interference fringes between light waves from a reference coherent light source (laser) and the light waves reflected from a subject. For microscale holography [31], holograms are typically captured with an interferometric setup equipped with a digital sensor. For macroscale holography [31], holograms are typically computationally generated from conventional Two Dimensional (2D) images, Computed Tomography (CT) scans, Magnetic Resonance Images (MRI), LF content, and point clouds. Holography is still maturing in terms of hardware technologies, mainly for macroscale holography applications [31]. However, there are some applications that have demonstrated the advantages of using this technology, for instance, for microscopically monitoring biological samples [32].

Figure 1.3 The 360º replay technology system [20] *based on an angularly arranged array of conventional cameras (left), and the application for FN demonstrated in the Super Bowl 50 (right)*

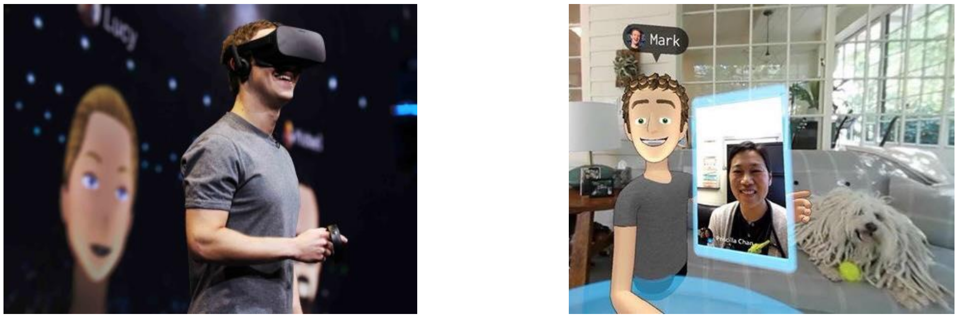


Figure 1.4 First demo for immersive social network application using virtual reality (presented by the Facebook founder Mark Zuckerberg at Oculus Connect 2016 [30]*)*

A critical factor for preventing these powerful new technologies from having the same fate as stereo technology in the consumer market is to create compelling and understandable use case scenarios in order to bring them to as many consumers' attention as possible. Moreover, a standard way to perform some of the aforementioned applications is also of the utmost importance in order to support interoperability between different (legacy and new) devices and services.

Therefore, recognizing the potential of these emerging technologies, as well as the new challenges that need to be overcome for successfully introducing them into the consumer market, novel standardization initiatives are also emerging, as will be seen in the next section.

1.2 Related Standardization Activities

Both Joint Photographic Experts Group (JPEG) and Moving Picture Experts Group (MPEG) standardization bodies have already started exploring these novel imaging applications to eventually support them in a standardized way. Notably, the JPEG committee has recently started the JPEG Pleno standardization initiative [33], and the MPEG group has started the third phase of Free-viewpoint Television (FTV), as well as new exploration work on Point Cloud and Light Field Compression.

A brief review of the developments in these standardization activities (at the time of writing this Thesis) is presented in the following.

1.2.1 JPEG Pleno Standardization

Realizing the emergence of new imaging technologies and representation modalities, JPEG Pleno was launched in 2015 [33] aiming at providing a standard framework to facilitate the capture, representation, and exchange of omnidirectional, depth-enhanced, point cloud, light field, and holographic imaging modalities [34].

To support the large amount of data involved in such systems, JPEG Pleno intends to define new tools for improved compression while providing advanced functionalities, for instance, for [31]: i) image and metadata manipulation; ii) spatial random access; iii) low latency and real time processing; iv) backward and forward compatibility with JPEG legacy formats; v) scalability; vi) privacy and security; and vii) parallel and distributed processing.

The first action towards the JPEG Pleno standardization has been identifying compelling use cases according to the industry, as well as defining the specific requirements for all the recognized imaging modalities [31]. Afterwards, the standardization process should start in different steps [34], in which the first one is going to address LF coding technologies. In this context, the JPEG committee has launched a new work item [35], and a final Call for Proposals on light field coding should be issued by the first quarter of 2017, according to the work plan available in [34].

1.2.2 MPEG FTV Standardization

MPEG FTV is a longtime ad-hoc group in MPEG, which has led the MPEG standardization initiatives that resulted in the H.262/MPEG-2 [36] Multiview Profile [37], H.264/MPEG-4 Advanced Video Coding (AVC) [38] standard extension for Multiview Video Coding (MVC) [39], as well as the recent HEVC extensions MV-HEVC and 3D-HEVC [40].

In 2013, the group started promoting the third phase of FTV standardization [41], aiming at exploring new data formats, compression and rendering technologies for Super Multiview (SMV) and FN applications [41]. In this context, the first action has been identifying use cases of interest for the industry and the specific requirements for SMV and FN systems [41]. For SMV systems, the requirement is to support more advanced LF displays that are emerging, which requires an ultra-dense and a very wide baseline set of linearly or angularly arranged views with horizontal parallax [41]. The challenge in this case is to deal with the massive amount of data required for these displays. For FN systems, the requirement is to enable users to view a scene by freely changing the viewpoints as is naturally experienced in the real world. For this, a sparse number of views with arbitrary positioning and wide baseline is considered, and the challenge is to support the generation of additional views at the decoder side so as to support real “fly through the scene” functionalities [41] (e.g., see Figure 1.3).

More recently, a call for evidence has been issued [42], aiming at gathering and evaluating proposals from companies and organizations for coding solutions, considering the set of identified SMV and FN application scenarios [41]. The main objective of this call has been to analyze if there is room for performance improvements (for the identified scenarios) compared to the coding, depth estimation and view synthesis algorithms used in the 3D-HEVC standard [42].

The evaluation of the responses to the call for evidence has been finalized and the complete results have just been summarized in [43]. The obtained results suggest that there are coding solutions that are better adapted to the considered SMV and FN application scenarios than those currently standardized. Regarding the Rate Distortion (RD) performance, some proposals present around 10 % of bitrate reduction compared to the 3D-HEVC for the same objective quality. Moreover, subjective evaluation also shows that, in some of the cases, statistically significant improvements can be observed [43] when compared to 3D-HEVC. Therefore, as mentioned in the call for evidence document [42], the following step might be to issue a call for proposals in coding technologies for SMV and FN.

1.2.3 MPEG Point Cloud Compression Standardization

Recognizing that there are prospective application scenarios out of the scope of the MPEG FTV ad-hoc group, MPEG has also started exploration into point cloud imaging applications.

For this, an ad-hoc group for point cloud compression has been established [44] (previously referred to as ad-hoc group for graphics compression [45]), in which use cases [46] and requirements [47] for point cloud compression have been discussed.

The requirements include the support for [47]: i) lossless and lossy point cloud compression (depending on the use cases defined in [46]); ii) progressive and scalable coding; iii) view-dependent decoding and spatial random access; iv) temporal random access; v) error resilience; vi) low complexity and low latency; and vii) parallel encoding and decoding.

The next action of the group, as stated in [48], will be to issue a call for proposals for point cloud compression by January 2017.

1.2.4 MPEG Exploration into Light Field Compression

More recently, an ad-hoc group for light field compression has been also established [44], focusing on the usage of LF imaging technology in applications that provide an increased sense of immersion, such as Six Degrees of Freedom (6DoF) VR applications [49].

The first action of this group has been to issue, by joint efforts of the JPEG committee and MPEG group, a report for digital representations of light/sound fields for immersive media applications [50]. The report identifies a prioritized list of use cases (according to the industry needs), provides a list of LF technologies for audiovisual applications, and defines a list of processing workflows for immersive applications scenarios [50].

Moreover, a call for LF test material has been just issued [49], which intends to collect natural and computer generated LF content, considering both LF content acquired using a dense array of multiple cameras, as well as acquired with a single camera equipped with an MLA [49]. The objective of this call is to enrich the existing MPEG content library so as to be able to conduct further experiments on LF compression [49].

1.3 Thesis Objectives and Original Contributions

This Thesis intends to address some of the aforementioned challenges and requirements for LF imaging applications, and then to contribute to the discussion in this field that has been recently initiated by the standardization bodies. More specifically, this Thesis aims at studying, developing and evaluating novel coding solutions for LF imaging applications that make use of the microlens-based LF imaging technology.

In this sense, three major objectives are defined for this Thesis, as listed below. For each of the defined objectives, a brief description of the related original contributions provided in this Thesis is also presented:

1) **Design of an Efficient LF Content Coding Solution** – One of the main challenges in LF imaging applications lies in the massive amount of data associated to this richer visual representation. Therefore, providing LF content coding tools that efficiently cope with this larger amount of data is a requirement of the utmost importance in order to deliver to the end-users LF content with convenient viewing resolution and with more powerful capabilities (e.g., in terms of content manipulation). This corresponds to the first objective of this Thesis.

   As a result of the used optical LF acquisition setup, the planar light intensity distribution in the LF content presents a repetitive structure of micro-images that may be exploited for achieving compression. Therefore, this Thesis proposes an efficient LF coding solution based on HEVC and using a non-local spatial prediction scheme, named Self-Similarity (SS) compensated prediction, to exploit this inherent redundancy that exists between micro-images in the LF content. For this, the proposed SS compensated prediction makes use of the generic concept of superimposed compensated prediction [51], in which one or more candidate predictor blocks can be estimated (according to appropriate criteria) from the same reference picture, and can be then used to predict the current block being coded. For this, the previously coded and reconstructed area of the current picture itself is seen as a reference frame, referred to as SS reference. Thus, similar to motion estimation, a block-based matching algorithm is used to estimate (inside a search window in the SS reference) the 'best' predictor block, in a Rate Distortion Optimization (RDO) sense [52], for the current block. This predictor block can be generated from a single candidate block [53–55], referred to as the uni-predicted SS candidate block, or from a combination of two different candidate blocks [56, 57], referred to as the bi-predicted SS candidate block. In the case of the bi-predicted SS candidate, the two candidate blocks are here jointly estimated by using the locally optimal rate-constrained algorithm proposed in [58] in order to further improve the RD coding performance [57]. As a result of this SS compensated prediction, one or two SS vectors are derived, which indicate the relative horizontal and vertical positions of the chosen candidate block(s) with respect to the position of the current block. To take advantage of the distinctive characteristics of these SS vectors, a novel SS vector prediction scheme is also proposed [59] in order to achieve further bit savings to the proposed LF coding solution. The main advantage of this LF coding solution based on SS compensated prediction is that it is not tuned for any particular optical acquisition setup since it does not require any explicit knowledge about it (e.g., microlens size, focal length, and distance of the microlenses to the image sensor). Moreover, the SS compensated prediction can be easily integrated to any hybrid coding architecture so as to extend it for LF video coding. This research work has resulted in the following publications:

- CONTI, Caroline, NUNES, Paulo and SOARES, Luis Ducla. New HEVC Prediction Modes for 3D Holoscopic Video Coding. In: *2012 19th IEEE International Conference on Image Processing*. Orlando, FL, US, September 2012. p. 1325–1328.
- AGGOUN, Amar, FATAH, Obaidulah Abdul, FERNANDEZ, Juan C J.C., CONTI, Caroline, NUNES, Paulo and SOARES, Luis Ducla. Acquisition, Processing and Coding of 3D Holoscopic Content for Immersive Video Systems. In: *2013 3DTV Vision Beyond Depth (3DTV-CON)*. Aberdeen, Scotland: IEEE, October 2013.
- CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. HEVC-Based 3D Holoscopic Video Coding using Self-Similarity Compensated Prediction. *Signal Processing: Image Communication*. March 2016. Vol. 42, p. 59–78.
- CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. 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, July 2016. p. 1–4.
- CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Light Field Image Coding with Jointly Estimated Self-Similarity Bi-Prediction. *Submitted to Signal Processing: Image Communication*.

At the time the work on this Thesis started, H.264/AVC represented the state-of-the-art on video coding technology. Hence, an LF coding solution based on H.264/AVC and using the SS compensated prediction has been also proposed, aiming at improving the RD performance of H.264/AVC for LF image [60, 61] and video [62] coding. In this case, only the uni-predicted SS candidate block is estimated in a search window inside the SS reference, and, consequently, a single SS vector is derived for each block. Experimental results using this preliminary version of the proposed LF coding solution has shown that substantial RD gains in LF image and LF video coding when compared with the conventional H.264/AVC are possible. Moreover, some preliminary experimental results in [54] have shown that, independently of the codec technology (namely, H.264/AVC or HEVC), the SS compensated prediction further improves the

RD performance relatively to the original version of the codec. However, as a result of the superior performance of HEVC compared to H.264/AVC, substantial higher RD gains are achieved by this HEVC-based LF coding solution when compared to the H.264/AVC-based solution proposed in [60, 61]. For this reason, the H.264/AVC-based LF coding solution is not further addressed in this Thesis. This preliminary research work resulted in the following publications:

- CONTI, Caroline, LINO, João, NUNES, Paulo, SOARES, Luís Ducla and CORREIA, Paulo Lobato. Improved Spatial Prediction for 3D Holoscopic Image and Video Coding. In: *19th European Signal Processing Conference (EUSIPCO 2011)*. Barcelona, Spain, August 2011. p. 378–382.
- CONTI, Caroline, LINO, Joao, NUNES, Paulo, SOARES, Luís Ducla and CORREIA, Paulo Lobato. Spatial Prediction Based on Self-Similarity Compensation for 3D Holoscopic Image and Video Coding. In: *2011 18th IEEE International Conference on Image Processing*. Brussels, Belgium, September 2011. p. 961–964.
- CONTI, Caroline, LINO, João, NUNES, Paulo and SOARES, Luís Ducla. Spatial and Temporal Prediction Scheme for 3D Holoscopic Video Coding based on H.264/AVC. In: *2012 19th International Packet Video Workshop (PV)*. Munich, German, May 2012. p. 143–148.
- CONTI, Caroline, SOARES, Luis D. and NUNES, Paulo. Influence of Self-Similarity on 3D Holoscopic Video Coding Performance. In: *Proc. of the 18th Brazilian symposium on Multimedia and the web - WebMedia ’12*. São Paulo, Brazil: ACM Press, October 2012. p. 131–134.

2) **Design of a Scalable LF Coding Solution to Provide Display Backward Compatibility** – In addition to the challenge of efficiently handling the massive 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. Dealing with this specific concern is an essential requirement for allowing faster deployment of new LF imaging application services in the consumer market. In this context, the second objective of this Thesis is to design an efficient scalable LF coding solution that aims at providing backward display compatibility.

   In order to achieve this, this Thesis proposes a display scalable LF coding architecture using a three-layer hierarchical approach, where each layer represents a different level of display compatibility [63]: i) 2D display (base layer); ii) stereo and multiview display (first enhancement layer); and iii) LF display (LF enhancement layer). This way, by decoding only the adequate subsets of the proposed scalable bitstream, it is possible to accommodate: i) end-users who want to have a conventional 2D visualization, since a simple 2D version of the LF content can be reconstructed by decoding only the base layer; ii) end-users who want to have conventional stereo or multiview visualization, since a stereo or multiview version of the same LF content is available by decoding only up to the first enhancement layer; as well as iii) end-users who want to have a more immersive and interactive visualization using more advanced LF display technologies (by decoding the entire scalable bitstream). Based on this hierarchical coding architecture, an LF enhancement codec is also proposed to efficiently encode the LF content in the highest hierarchical layer. For this, the SS compensated prediction, which has been proposed in the previous contribution, is proposed to be combined with a novel inter-layer prediction scheme for improving the RD coding performance compared to independent compression of the three different layers (i.e., the simulcast case) [64–66]. The proposed inter-layer prediction mechanism aims at exploiting the existing redundancy between the stereo/multiview and the LF content. To accomplish this, a prediction picture is built and is then used as a new reference frame in an inter-layer compensated prediction scheme. The publications related to this contribution are:

- CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Inter-Layer Prediction Scheme for Scalable 3-D Holoscopic Video Coding. *IEEE Signal Processing Letters*. August 2013. Vol. 20, no. 8, p. 819–822.
- CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Using Self-Similarity Compensation for Improving Inter-Layer Prediction in Scalable 3D Holoscopic Video Coding. In: *Proc. SPIE 8856 Applications of Digital Image Processing XXXVI*. San Diego, CA, US, September 2013. p. 88561K.
- CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. Display Scalable 3D Holoscopic Video Coding. *IEEE Communications Society MMTC E-Letter*. May 2014. Vol. 9, no. 3, p. 12–15.
- CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. 3D Holoscopic Video Representation and Coding Technology. In: KONDOZ, Ahmet and DAGIUKLAS, Tasos (eds.), *Novel 3D Media Technologies*. Springer New York, 2015. p. 71–96.

3) **Design of a Scalable LF Coding Solution with Support for Flexible Interactive Manipulation Functionalities** – Among the exciting new interactive functionalities that future LF imaging applications support are the possibility for post-production refocusing, changing depth of field, and changing viewing perspective. This means that, for instance, the end-user can receive a captured LF and interactively adjust the plane of focus and depth of field in 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, the third objective of this Thesis is to design a scalable coding architecture that supports different levels of

interactivity, being able to accommodate in a single compressed bitstream a variety of sub-bitstreams appropriate for end-users with different preferences/requirements and various application scenarios. The targeted users range from the end-user who wants to have a simple 2D version of the LF content without actively interacting with it; to the end-user who wants full immersive and interactive LF visualization.

To achieve this, this Thesis proposes a novel scalability concept, named Field Of View (FOV) scalability, as well as a new scalable coding solution with FOV scalability [67]. Taking advantage of the particular radiance distribution in the LF content, the FOV scalability progressively supports richer forms of the same LF content by hierarchically organizing the angular information of the captured LF content. More specifically, the base layer contains a subset of the LF data with narrower FOV, which can be used to render a 2D version of the content with very limited interactive 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 which are adaptable to different levels of user interaction, allowing increasing degrees of freedom in content manipulation at each higher layer. In addition to the FOV scalable LF coding architecture, two novel inter-layer prediction schemes are also proposed in this Thesis [67] to efficiently encode the LF data in each enhancement layer, namely: i) a direct prediction scheme based on exemplar texture samples from lower layers; and ii) an inter-layer compensated prediction scheme using a reference picture that is constructed relying on a patch-based algorithm for texture synthesis. In the exemplar-based direct prediction, a set of available samples from a previous layer is used as exemplar samples for estimating a good prediction block. Therefore, there is no need to send displacement vectors, since these vectors can also be derived at decoder side. In the patch-based inter-layer reference picture construction, samples from the previous layers are used to estimate the unknown areas of the inter-layer reference picture, based on an optimization algorithm adapted for the inherent constraints of the LF content. This research work has been submitted to the following journal:

- CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. Light Field Coding with Field of View Scalability for Flexible Interaction. *Submitted to IEEE Transactions on Multimedia*.

1.4 Other Contributions during this Thesis

In addition to the abovementioned contributions, some other work has been done during this Thesis also in the field of LF content coding, which is briefly described in this section. These contributions are not further addressed in this Thesis since they are not directly related to the defined objectives, or since it has been done in joint collaboration with other research teams. These are:

1) **Impact of Packet Losses in Scalable LF Coding** – To effectively transmit LF content over error-prone networks, e.g., wireless networks or the Internet, error resilience techniques are required to mitigate the impact of data impairments in the end-user quality perception. Therefore, it is essential to deeply understand the impact of packet losses in terms of the decoding video quality for the specific case of LF content, notably when a scalable LF coding approach is used [68]. In this context, this work aims at studying the impact of packet losses when using the three-layered display scalable LF coding architecture proposed in this Thesis. For this, a simple error concealment algorithm is used, which makes use of inter-layer redundancy between multiview and LF content as well as the inherent correlation of the LF content to estimate lost data. Furthermore, a study of the influence of 2D views generation parameters used in lower layers on the performance of the used error concealment algorithm is also done. The results of this study have been published in the following conference:

   - CONTI, Caroline, NUNES, Paulo and DUCLA SOARES, Luís. Impact of Packet Losses in Scalable 3D Holoscopic Video Coding. In: *Proc. SPIE Optics, Photonics, e Digital Technologies for Multimedia Applications III*. Brussels, Belgium, May 2014. p. 91380E.

2) **LF Video Coding Using Geometry-Based Disparity Compensation** – This work studies and evaluates two HEVC-based coding solutions for efficient compression of LF content acquired using a dense array of horizontally arranged cameras. These two coding schemes aim at exploiting the 3D geometry-based disparity information known from the acquisition process so as to replace the block-based disparity estimation [69]. In the first scheme, the disparity map of each view is used to directly derive the vectors for compensation, and in the second scheme these disparity vectors (for all views) are calculated (for non-occluded areas) from the disparity map of the base view. The results of this research work have been published in the following conference:

   - CONTI, Caroline, KOVACS, Peter Tamas, BALOGH, Tibor, NUNES, Paulo and SOARES, Luis Ducla. Light-Field Video Coding Using Geometry-Based Disparity Compensation. In: *2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. Budapest, Hungary, July 2014.

3) **LF Content Coding Using Locally Linear Embedding (LLE) Prediction** – An alternative non-local spatial prediction, known as LLE prediction, is studied in this work to be used in an HEVC-based LF image coding solution. The basic idea of the LLE prediction is to search, inside a causal search window in the LF image, for an

optimized linear combination of *k*-Nearest Neighbor (*k*-NN) texture patches that best approximate (in terms of least-squares optimization) known sample values in a previously coded neighborhood of the current block (known as template). Then, the same combination of coefficients is used to approximate the unknown samples in the current block [70]. The results of this work have been published in the following conference, having been nominated as one of the four candidates to the EUSIPCO 2014 Best Student Paper Awards:

- LUCAS, Luís F. R., CONTI, Caroline, NUNES, Paulo, SOARES, Luís Ducla, RODRIGUES, Nuno M. M., PAGLIARI, Carla L., DA SILVA, Eduardo A.B. and DE FARIA, Sergio M. M. 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, September 2014. p. 11–15.

1.5 Thesis Outline

This Thesis tackles the design of efficient coding solutions for LF imaging applications. In order to facilitate the reading, the organization of this Thesis is illustrated in Figure 1.5.

In this current chapter, the context and motivation for this work are presented, along with the definition of the Thesis objectives and a summary of the main original contributions.

Chapter 2 briefly reviews the principles of LF imaging technology, as well as the basics and recent developments related to LF acquisition, LF representation, LF rendering, and LF display.

Chapter 3 focuses on the coding requirements and reviews the most relevant LF image and LF video coding approaches proposed in the literature. In addition, an overview of 2D and 3D coding standards that are relevant for this Thesis is also presented.

After this initial overview, the original contributions of this Thesis related to LF content coding are addressed. In this context, Chapter 4 tackles the first objective of this Thesis (see Section 1.3) and proposes an efficient LF coding solution based on HEVC and using the concept of SS compensated prediction.

Following this, Chapter 5 deals with the requirement for providing backward compatibility between LF and legacy devices, which corresponds to the second objective of this Thesis (see Section 1.3). In this case, a display scalable LF coding solution is proposed as well as a novel inter-layer prediction scheme that is then combined with the SS compensated prediction.

Chapter 6 tackles the third objective of this Thesis (see Section 1.3), which is to provide LF coding solutions that support flexible interactive functionalities. In this case, a novel

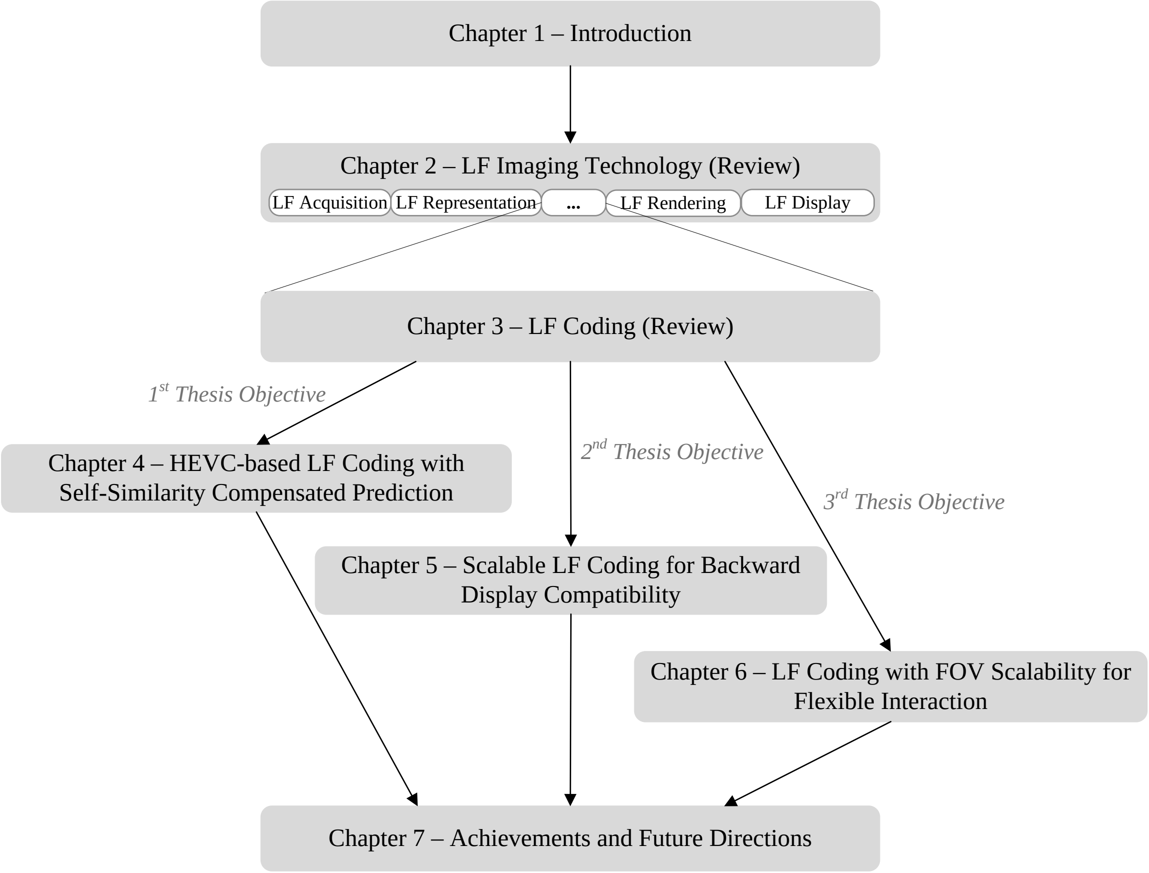

Chapter 1 – Introduction
Chapter 2 – LF Imaging Technology (Review)
LF Acquisition
LF Representation
...
LF Rendering
LF Display
Chapter 3 – LF Coding (Review)
1st Thesis Objective
2nd Thesis Objective
3rd Thesis Objective
Chapter 4 – HEVC-based LF Coding with Self-Similarity Compensated Prediction
Chapter 5 – Scalable LF Coding for Backward Display Compatibility
Chapter 6 – LF Coding with FOV Scalability for Flexible Interaction
Chapter 7 – Achievements and Future Directions

Figure 1.5 Organization of this Thesis

scalability concept, the FOV scalability, as well as a FOV scalable coding architecture are proposed, along with two novel inter-layer exemplar-based prediction schemes.

Finally, Chapter 7 concludes this Thesis by discussing the achievements and by identifying some directions for future work.

After this closing chapter, Appendix A presents a description of all LF images and LF video sequences that are used throughout the Thesis, as well as a description of the process used in preparation for the coding stage. Following this, Appendix B presents some additional experimental results for the LF coding solution proposed in Chapter 4, considering LF images captured using a Lytro Illum LF camera.

Chapter 2

Light Field Imaging Technology

Before concentrating the discussion on LF coding solutions in the following chapters, this chapter reviews the principles of LF imaging technology and provides a comprehensive overview of the main stages that are essential for efficiently delivering LF content to the end-users.

For this, this chapter starts by defining the concept of light fields in Section 2.1. Following this, a conceptual analysis of the basic LF imaging processing chain [35, 71] is presented, comprising the following functional modules, as illustrated in Figure 2.1:

- **LF Acquisition/Creation** – The first step of the LF processing chain is, naturally, the LF content generation, which can be done through an optical setup, i.e., acquired by an LF camera or an array of common cameras or computationally created, i.e., through appropriate modeling and rendering of the visual scene and the acquisition setup. Therefore, Section 2.2 reviews the principles of LF imaging acquisition, with a particular emphasis on the LF imaging technology that makes use of an MLA in the optical setup.

- **LF Representation** – The LF data acquired in the previous stage may or may not be converted to a representation format that is different than the acquired format. In this context, Section 2.3 presents a brief review of representation formats that have been proposed in the literature.

- **LF Encoding/Decoding** – Considering the huge amount of data associated to LF imaging systems, efficient LF coding/decoding solutions become of paramount importance. Since this is the main focus of this Thesis, a more detailed overview of these issues is presented in Chapter 3.

- **LF Rendering** – Rendering the decoded LF content becomes also an important issue, especially to allow adequate visualization of the decoded LF content in conventional 2D and 3D displays. Therefore, Section 2.4 addresses this issue and reviews some rendering algorithms that are relevant in the context of this Thesis.

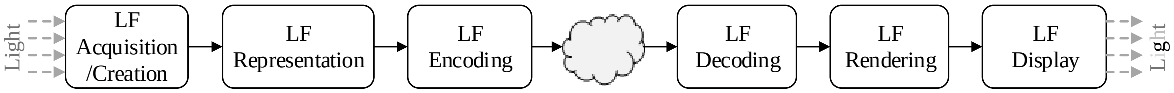


Figure 2.1 LF imaging processing flow [35, 71]

- **LF Display** – To take full advantage of the richer visual information of the acquired LF content, new and more immersive display devices are also needed. For this reason, display technologies have been also evolving in the recent years, and Section 2.5 overviews the recent developments related to this.

In addition to this, the display technology along with the rendering capabilities will also determine what should be expected in terms of the user experience in LF imaging applications. This fact brings up another challenging issue, which is designing appropriate objective and subjective metrics for LF quality evaluation. This issue is then briefly reviewed in Section 2.6.

2.1 Light Fields

Before starting the discussion on the processing chain presented in Figure 2.1, it is important to describe the fuel of this entire chain, i.e., the visible light, as well as to review the concept of light fields.

The term light field was firstly adopted by Gershun in 1936 for analytically describing the "*light beams that propagate in a straight line through a homogeneous medium and which is the carrier of radiant energy in a space*" [72]. Generally, the concept of light fields comes from the necessity of describing and also replicating the complete/full/whole visible light information in a given surrounding as accurately as possible. Actually, this notion had been exploited earlier by the polymath Da Vinci (1452-1519) when he suggested the existence of the pyramids of sight [73], as well as by the physics Nobel laureate Lippmann, in 1908, when he introduced the concept of integral photography [8]. More recently, the terms holoscopic (from the Greek *holōs* (whole) + *optikos* (vision)) [74], plenoptic (from the Latin *plenus* (full) + *opticus* (vision)) [75], and lumigraph [76] have been also adopted almost as synonymous.

In early 1990s, with the popularity growth of computer vision and computer graphics, the problem of geometrically describing the visible space gained a major importance. Notably, Adelson and Bergen proposed, in [75], to define the total distribution of light as a Seven Dimensional (7D) function, which models the light rays at every possible location in space (x,y,z), toward every possible direction (θ,φ), over any range of wavelengths (λ), and at any time (t) – referred to as the plenoptic function (see Figure 2.2):

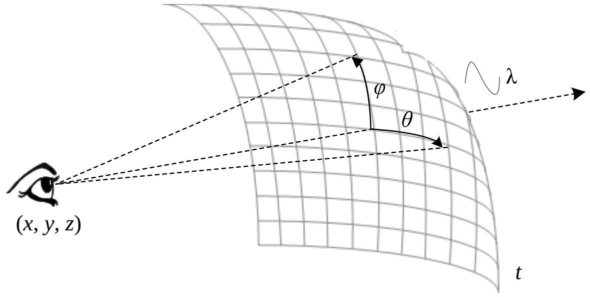


Figure 2.2 The 7D plenoptic function

$$P(x, y, z, \theta, \varphi, \lambda, t) \tag{2.1}$$

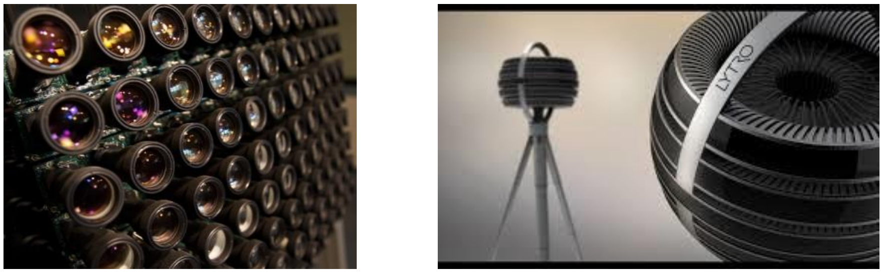
Thus, with this definition, it is possible to model different imaging systems, including the Human eye [75], as samplings from the 7D plenoptic function in (2.1). In fact, due to the enormous amount of data that would be required for sampling using a 7D representation, it is necessary to make reasonable assumptions to reduce the dimensionality of the plenoptic function and to appropriately sample it.

Therefore, Levoy and Hanrahan proposed, in [77], to use the following three assumptions to reduce from the 7D plenoptic function to the Four Dimensional (4D) light field function [77] (also known as lumigraph [76]):

1) **Static Scene** – Assuming the scene is static, the plenoptic function can be then reduced to $P(x,y,z,\theta,\varphi,\lambda)$.

2) **Constant Radiance along its Path (in a Free-Space)** – With the assumption that the air is truly transparent and the light ray is transmitted in a free-space (i.e., region free of occluders [77]), the plenoptic function can be then represented by its values along an arbitrary selected surface surrounding the scene (see Figure 2.2). Hence, the radiance of any light ray in the space can be always obtained by tracing it back to this selected surface. This assumption allows reducing the plenoptic function to $P(x,y,\theta,\varphi,\lambda)$.

3) **Trichromatic Human Vision System (HVS)** – The Human eye has three types of photosensitive cells (known as cones) in the retina for the perception of colored light. Each of these cone types has its maximum sensitivity in a different wavelength, which corresponds to the primary colors Red (R), Green (G), and Blue (B). Therefore, it is possible to restrict to the HVS and to reduce the wavelength dimension in (2.1) by assuming three different plenoptic functions (one for each R, G, and B components). Finally, for each color component, a 4D light field function is defined as $P(x,y,\theta,\varphi)$.



(a) *(b)*

Figure 2.3 Examples of light field capturing systems based on multi-camera arrays: (a) The Stanford multi-camera array [78], *in which conventional cameras are arranged regularly in a linear array with full parallax; and (b) The Lytro Immerge system of cameras* [12], *in which a globe with rigs of multiple cameras is used to capture real LF content with 6DoF*

In the context of this Thesis, it will suffice to describe the complete visible light by this 4D light field function[1] [77]. This means that light is here understood as a scalar radiance (one for each color component R, G, and B) traveling along straight lines (rays) with different directions.

2.2 LF Acquisition

After having defined the concept of light fields in the previous section, it is now possible to tackle the basics of the processing chain in Figure 2.1. Notably, this section addresses the first LF acquisition stage.

With the target of increasing immersion, more advanced imaging technologies are emerging that allow capturing richer forms of visual data and representing the scene by the aforementioned 4D light field. Among these technologies, two main groups may be listed:

- **LF Imaging with Multi-Camera Array** – In this case, a dense number of views (also known as SMV) with single direction or full parallax is captured using a multiple-camera array (see Figure 2.3) in a linear [78], circular [12], or even arbitrary arrangement.
- **LF Imaging with Single-Tier Camera and MLA (Microlens-Based LF)** – In this case, the integral photography concept proposed by Lippman [8] (also known as integral [8, 9], holoscopic [5], and plenoptic [10] imaging) is adopted, in which LF content with full parallax can be acquired by using a single-tier sensor camera equipped with an MLA.

[1] Notice that, although the LF coding solutions proposed in this Thesis are developed targeting still LF images, they can be easily extended for LF video coding. In this case, light can be then described as a 5D plenoptic function $P(x,y,\theta,\varphi,t)$.

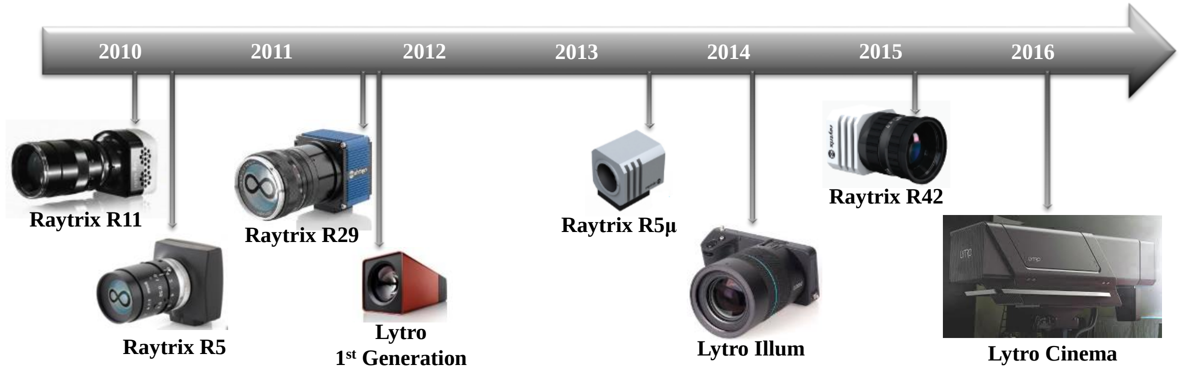


Figure 2.4 Timeline of the most relevant LF cameras [11, 12] *that have been announced in the market*

It should be notice that, this Thesis will be completely devoted to the microlens-based LF (also known as lenslet LF) imaging technology and, for this reason, it will be simply referred to as LF imaging from hereinafter.

Following the recent developments in optical and sensor manufacturing, LF cameras have recently achieved a major breakthrough, becoming increasingly available in the consumer market. Figure 2.4 depicts a timeline of the most relevant LF cameras from Raytrix [11] and Lytro [12] family since the first commercially available LF camera (i.e., Raytrix R11 [11]) has been announced. Moreover, among the advantages of using an LF camera for acquiring LF content, one can cite:

- **Synchronization by Design** – Since a single tier camera system is used, there is no need for complex synchronization procedures between several cameras, such as to perfectly match their orientation, focal length, and exposure.
- **Highly Dense LF Content** – The sensor underneath the MLA is able to record the scene from hundreds or thousands of different viewpoint positions with smooth horizontal and vertical parallax.

As illustrated in Figure 2.5, an LF camera basically comprises a main lens, and an image sensor overlaid with a MLA. However, as discussed in [79, 80], from these basic elements, two LF camera setups can be further derived, namely: i) the traditional LF camera setup; and ii) the focused LF camera setup. The characteristics of these two setups are discussed as follows.

2.2.1 The Traditional LF camera setup

In a traditional LF camera, also known as plenoptic camera 1.0 [81], (see Figure 2.5a), the main lens is focused on the microlens plane while the microlenses are focused at infinity. This is illustrated in Figure 2.5a where the sensor is placed at the MLA focal length, f and since microlenses have, usually, a much smaller focal length than the main lens, it is

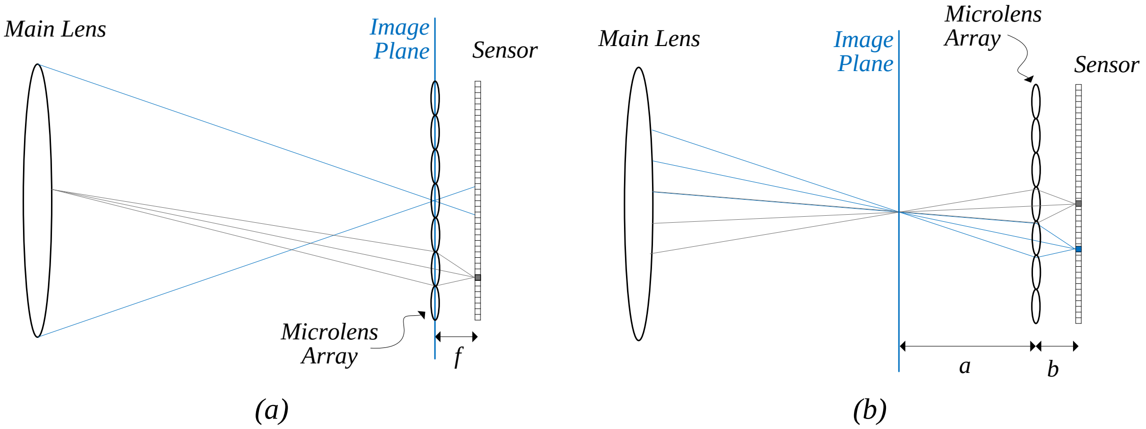


Figure 2.5 Different setups for the LF camera: (a) The traditional LF camera setup; (b) The generalized focused LF camera setup

reasonable to admit that the main lens is at the microlenses optical infinity. This means that the light rays coming from the main lens will be thus focused on the sensor, as illustrated by the gray cone of rays in Figure 2.5a. In addition, it is illustrated by the blue rays (Figure 2.5a) that all the light rays coming from a depth plane that is in sharp focus for the main lens (i.e., converging to the MLA plane) will, consequently, intersect at the MLA and then diverge until they reach 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[2]); 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 light field and organize it in a conventional 2D image, known as the (raw) LF image. An illustrative example of an LF image captured using this camera setup can be seen in Figure 2.6.

Moreover, examples of LF cameras using this setup are the Lytro LF cameras [12] (see Figure 2.4).

2.2.2 The Focused LF camera setup

As discussed in [79, 80, 82], the traditional LF camera (Figure 2.5a) can be generalized to the alternative camera setup shown in Figure 2.5b, which is referred to as focused camera setup [80] (also known as plenoptic camera 2.0 [81]). Examples of commercially available focused LF cameras are the Raytrix LF cameras [11] (see Figure 2.4).

[2] 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, i.e., R, G, and B.

In the focused LF camera, the microlenses are no longer focused at infinity, but they are rather focused on the main lens image plane, as illustrated in Figure 2.5b. Hence, the name "focused" derives from the fact that each microlens is now focused on the same subject as the main lens [80]. In other words, each microlens forms a relay system with the main lens, which satisfies the lens equation in (2.2), i.e.,

$$\frac{1}{f}=\frac{1}{a}+\frac{1}{b} \tag{2.2}$$

where a is the distance between the main lens image plane and the MLA, and b is the distance between the MLA and the image sensor. An illustrative example of an LF image captured using this setup can be seen in Figure 2.6

In practice, varying between these two LF camera setups will only change the balance between providing larger directional or spatial resolution in the captured LF image (respectively, using a traditional or a focused LF camera) [79, 83]. In fact, the setup shown in Figure 2.5b, can be used to generalize this tradeoff between maximal directional and spatial resolution in an LF camera. For instance, when $b \rightarrow f$ and $a \rightarrow \infty$, the setup in Figure 2.5b will correspond to the traditional LF camera (Figure 2.5a) in which maximal directional resolution is allowed (notice that (2.2) still holds). On the other hand, in the case where $b \rightarrow 0$ and $a \rightarrow 0$, the setup in Figure 2.5b will correspond to a conventional 2D camera in which maximum spatial resolution is achieved in the sensor at the expense of no directional information.

Moreover, the main lens can be focused on a plane in front of the MLA plane (as seen in Figure 2.5b) or behind it. Depending on this choice, the image captured behind each microlens may be real (for the case in Figure 2.5b) or virtual (otherwise). For more details about the difference between these two options, the reader should refer to [84].

2.3 LF Representation

A key issue for successful LF imaging applications is the choice of a convenient representation for the LF raw data acquired in the previous stage, given a certain set of applications requirements. If, for example, high compression efficiency is a dominant requirement, then this decision shall be made prioritizing a coding perspective, which means that an efficient coded representation should be at the forefront.

In this context, this section briefly describes some relevant LF representation formats that have been proposed in the literature in the context of LF content coding. However, it should be noticed that the analysis and discussion of specific LF coded representations performance will be done in the following chapters.

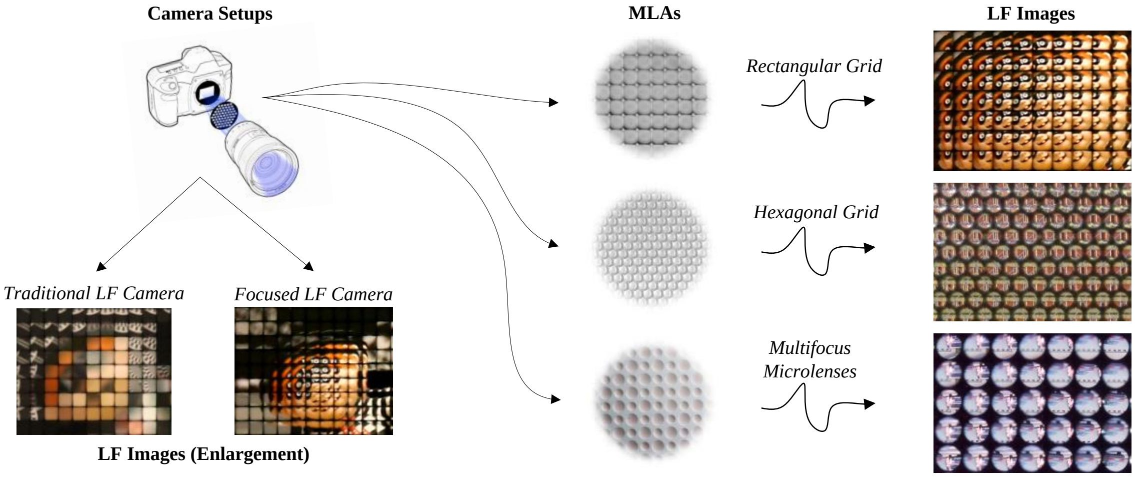


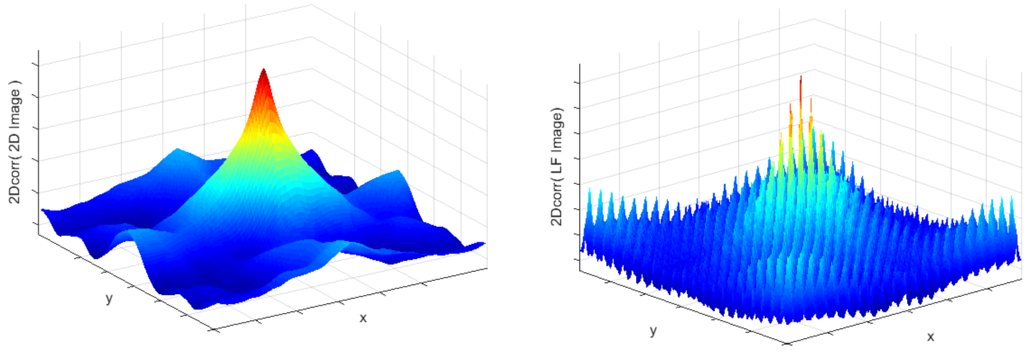
Figure 2.6 Examples of different LF camera setups and MLA structures. The micro-images structure is related to the chosen optical setup

2.3.1 Micro-Image-Based Representations

As a result of the used optical acquisition setup, where the MLA is overlaid with the image sensor, the planar light intensity distribution representing the LF image corresponds to a 2D grid of micro-images, as illustrated in Figure 2.6. As can be seen, each Micro-Image (MI) captures a low resolution portion of the scene at slightly different perspectives. Moreover, several packing schemes, shapes and sizes of microlenses are possible in the array (see Figure 2.6), and the structure of these MIs is a consequence of the chosen MLA. In addition to this, the MI characteristics may also change depending on the chosen LF camera setup (see Figure 2.6). For instance, for a traditional LF camera (Figure 2.5a), an MI is an image focused on the main lens (i.e., it is a picture of the back of the main lens [79]); while for a focused LF camera (Figure 2.5b), an MI is (a low resolution portion of) the image of the main lens that is relayed through the microlens.

Regarding MI-based representations for the LF data, the following two formats have been proposed in the literature, more specifically, in the context of LF content coding:

- **(Raw) MI-Based 2D Format** – This format corresponds to the simplest case, in which the LF acquisition format is adopted to be the representation format and, consequently, no conversion and/or further processing is required. Hence, the LF data is represented as a 2D image comprising a grid of MIs.
- **MI-Based Multiview 3D Format** – This 3D format here includes generically any format that supports the representation of two or more slightly different (displaced) views. In the case of an MI-based multiview 3D format, the LF image needs to be first split into its multiple MIs, which are then represented as multiple views. Hence, further calibration/processing is required in this case, for instance, to compute the MI centers,



(a) *(b)*

Figure 2.7 Comparison of the autocorrelation function for: (a) A 2D view rendered from the LF test image Plane and Toy (frame 123), as seen in Figure A.8b (right); and (b) The LF image Plane and Toy (frame 123), as seen in Figure A.8b (left)

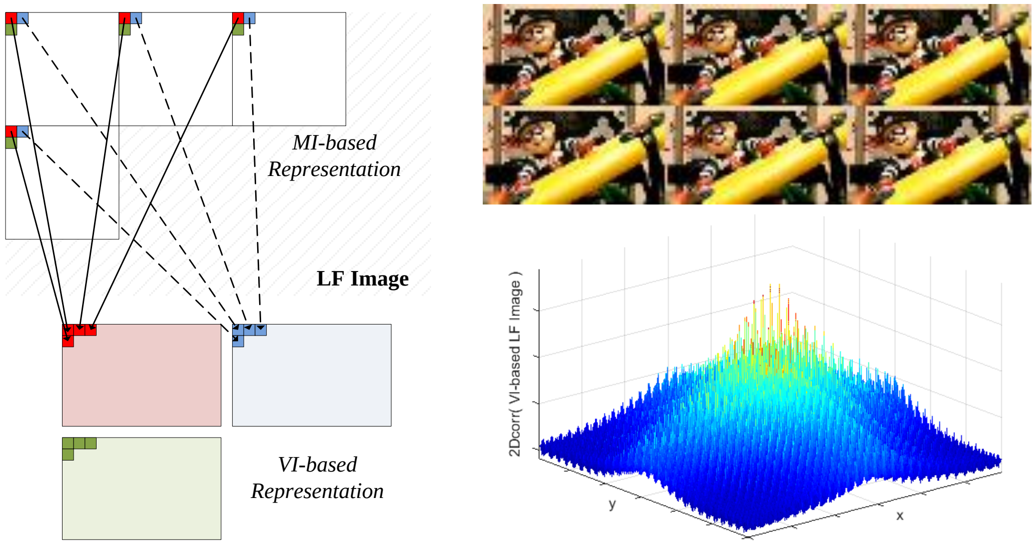
and to compensate any potential optical/geometrical distortions that may result in MIs with different sizes and with non-integer resolutions. Examples of calibration/processing algorithms for LF images can be found in [13, 85].

Analyzing these MI-based representations from a coding point of view, it is observed that, independently of the camera setup or MLA used, the LF content presents some inherent spatial correlations, as illustrated by the autocorrelation function in Figure 2.7b. Notably, it can be seen that the pixel correlation in an LF image is not as smooth as in conventional 2D images (see Figure 2.7a). Differently, a regular structure of spikes is evidenced in the autocorrelation function in Figure 2.7b, in which the constant distance between these regular spikes corresponds to the MI spacing in the array. Moreover, as is commonly observed in 2D images (see Figure 2.7a), pixels inside each MI are also significantly correlated within a local neighborhood (see Figure 2.7b).

2.3.2 Viewpoint Image-Based Representations

Using the knowledge of the exact LF optical setup (e.g., MI coordinates and sizes), a Viewpoint Image (VI) can be constructed, as illustrated in Figure 2.8a, by extracting one pixel with the same relative position from all MIs. Hence, several low resolution VIs can be extracted at different positions relative to the MI center, as illustrated in Figure 2.8b (top).

For an LF image captured using the traditional LF camera setup (see Figure 2.5a), each VI represents an orthographic projection of the captured scene [79]. On the other hand, for an LF image captured using a focused LF camera setup (see Figure 2.5b), a VI can be interpreted as a low resolution perspective of the captured scene that is focused at infinite [10].



(a) (b)

Figure 2.8 LF data representation based on VIs: (a) Converting from an MI-based representation to a VI-based representation; and (b) 3×2 VIs extracted from the LF image Plane and Toy (frame 123) seen in Figure A.8b (top), and the autocorrelation function (bottom) of the complete 2D grid of VIs

It is worth noticing that, before the conversion from an MI-based to a VI-based data organization (Figure 2.8a), additional calibration/processing may be required to compute the MI centers, to deal with MIs with slightly different sizes and with non-integer resolutions, and to align the MI grid to the pixel grid (see, for example, [13, 85]).

Therefore, similarly to the MI-based representation, the following two VI-based representation formats are worth mentioning:

- **VI-Based 2D Format** – This 2D format is based on VIs instead of MIs. For this, all extracted VIs are organized in a 2D grid to be represented as a 2D content (see Figure 2.8b, top). This new 2D image is referred here to as VI-based LF image. As can be seen in Figure 2.8b (bottom), there is also a significant correlation between VIs in a neighborhood.
- **VI-Based Multiview 3D Format** – In this case, each VI is seen as a view, and the LF data is then represented by its multiple VIs as 3D content.

2.3.3 Other LF Data Representations

Other LF data representations can be also designed by extracting the depth/disparity information as well as information on the intensity and direction of each particular light ray (i.e., the ray-space) from the captured LF image.

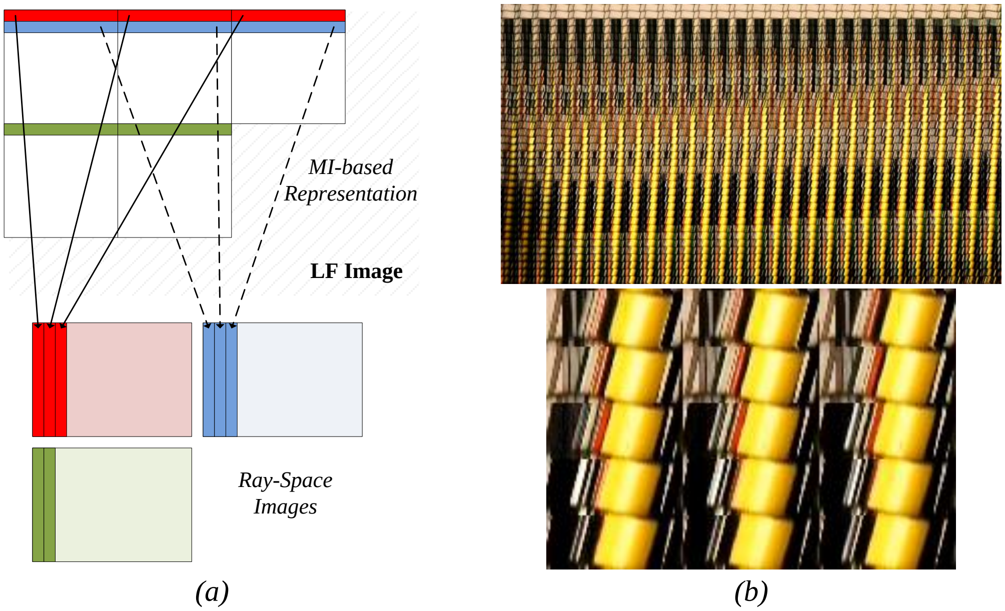


Figure 2.9 LF data representation based on ray-space images: (a) Extracting ray-space images from an MI-based representation; and (b) All ray-space images (top) extracted from the LF test image Plane and Toy (frame 123) (see Figure A.8b), and a 3×5 ray-space images in detail (bottom)

Therefore, the following two alternative representations formats are also relevant:

- **Multiview plus Depth/Disparity 3D Format** – In this case, the LF data is represented by a fixed number of views together with the associated depth/disparity information. Generically, the set of views may comprise MIs/VIs that are extracted from the LF data, or views with higher resolution that are rendered using a specific rendering algorithm (as discussed in Section 2.4). As depth/disparity information can be used for synthesizing views at the decoder side, the amount of views that needs to be coded and transmitted in the processing chain may be reduced. However, the accuracy of the depth/disparity estimated from the captured LF data may be a problem for having high quality synthesized views [86, 87]. Trying to deal with this problem, various depth/disparity estimation methods have been recently proposed in the literature (e.g., in [86–90]).

- **Ray-Space-Based Format** – In this case, the LF data can be decomposed according to its light ray distribution. For this, the Epipolar Plane Image (EPI) technique [91] can be used to generate the so-called ray-space images [92, 93]. Each ray-space image can be then interpreted as a 2D cut through the captured 4D light field (for instance, by fixing dimensions y and φ and varying x and θ). As an illustrative example, Figure 2.9 shows ray-space images that were capture by stacking together MIs in the same row (which corresponds to fixing the dimension y) and, then, taking a slice from these MIs in a particular horizontal plane (which corresponds to fixing the direction φ), as illustrated in Figure 2.9a. A prospective characteristic of this ray-space based representation is that the depth/disparity of the objects can be estimated from the slope of the lines that can be seen in Figure 2.9b. Examples of this usage for LF content acquired using a multi-camera array can be found in [94–96].

2.4 LF Rendering

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

In this context, one important requirement for the design of LF rendering algorithms is to offer the best end-user experience targeting a specific display technology [71]. Moreover, this requirement may eventually consider not only the visual quality of the rendered content, but also the level of user interaction that is enabled.

In this sense, among the advantages of the LF imaging technology is the ability to open new degrees of freedom in terms of content rendering, supporting functionalities not straightforwardly available using conventional single-tier sensor systems, such as:

- **Changing Perspective** – By simply re-tracing the light rays that are captured in the LF image, it is possible to render 2D images with different viewing angles.
- **Changing Focus (Refocusing)** – Briefly, rendering with refocusing can be seen as virtually translating the image plane of the LF camera (see Figure 2.5) to a different plane in front or behind it.
- **Depth of Field Control** – Increasing or decreasing the depth of field in the rendering simply means to define greater or smaller (discrete) number of depth planes to be in focus simultaneously.

As will be further discussed in Chapter 6, it is important to notice that the aforementioned rendering functionalities are constrained (e.g., in the number of effective different viewing angles or depth planes that are available for rendering) by the main lens aperture size, since it limits the amount of light rays that are admitted through the camera optical system.

Several LF rendering algorithms have been proposed in the literature, mainly in the context of richer 2D image capturing systems [10, 79, 97–100]. In this context, two of these algorithms, which have been proposed in [10] for focused LF cameras, are briefly reviewed in this section: i) Basic Rendering; and ii) Weighted Blending. These rendering algorithms are of direct relevance for this Thesis, mainly since they have been adopted for experimentally evaluating the performance of the LF coding solutions that are here proposed. More details about these algorithms can be found in [10].

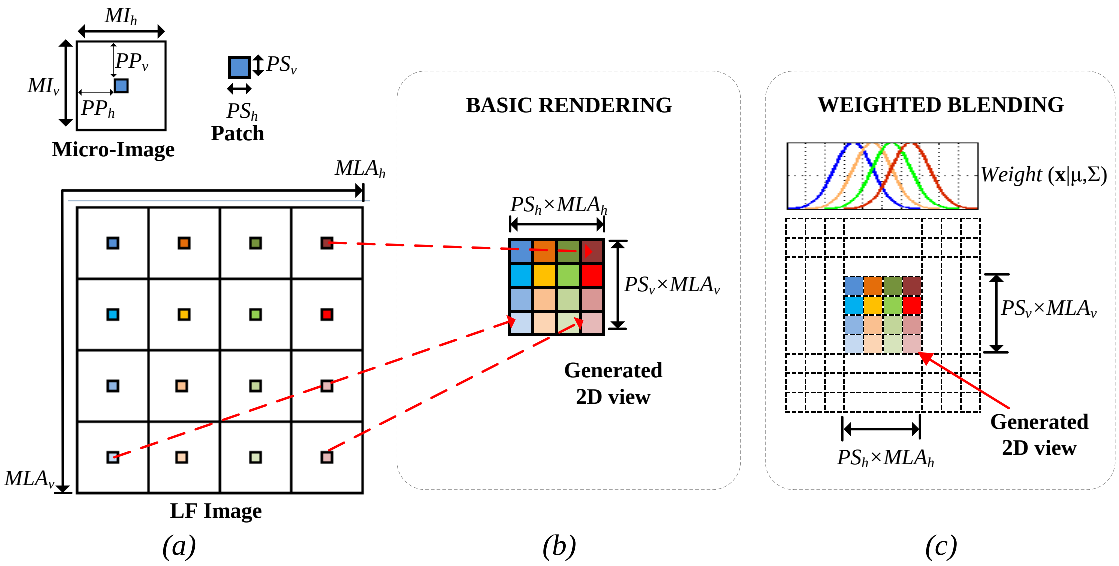


Figure 2.10 Rendering algorithms proposed in [10] *for focused LF camera setups: (a) Patch location in the input LF image; (b) Basic Rendering algorithm; and (c) Weighted Blending algorithm. From: Conti, Nunes, and Soares, 2013* [63]

2.4.1 Basic Rendering Algorithm

The idea of the Basic Rendering algorithm is that, since each MI can be seen as a partial low resolution version of the captured scene, a 2D view of the scene can be rendered by selecting and stitching together suitable sets of samples (patches) from each MI.

Therefore, the input for this algorithm is an LF image (see Figure 2.10a [63]) with an $MLA_h \times MLA_v$ array of MIs, where each MI has a resolution of $MI_h \times MI_v$. In this LF image, a patch of $PS_h \times PS_v$ pixels is extracted from each MI. These patches are then stitched together, as illustrated in Figure 2.10b. As a result, the output is a rendered 2D view of the scene with a resolution of $(PS_h \times MLA_h) \times (PS_v \times MLA_v)$.

There are the two main parameters that control this algorithm:

- **Patch Size** – Since an MI is captured by the corresponding microlens in perspective projection geometry, placing two objects of the same (real) size at different depths, i.e., one closer and the other farther from the MLA, will result in those objects appearing with greater or smaller size in pixels in the various MIs. Thus, for a given object to appear sharp in the generated 2D view image, a different patch size (larger and smaller, respectively, for the closer and for the farther object) needs to be selected from each MI. This fact is explained in more detail in [10], where it is also shown that it is possible to control the refocusing functionality in the LF rendering by simply choosing a suitable patch size.

- **Patch Position** – By varying the relative position of the patch inside the MI (given by PP_h and PP_v in Figure 2.10), it is possible to generate 2D views with different horizontal and vertical viewing angles (i.e., to change the rendering perspective). Therefore, it is also possible to render 3D multiview content from the LF data by selecting two (stereo) or multiple (multiview) appropriate viewing perspectives.

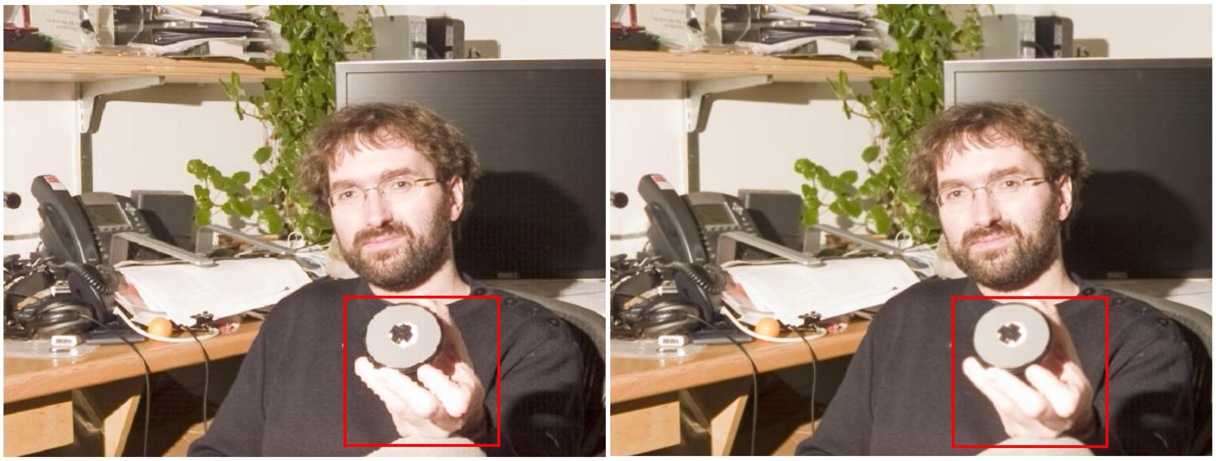
2.4.2 Weighted Blending Algorithm

To avoid some blocking artifacts due to a non-perfect match between the stitched patches with fixed sizes in the Basic Rendering algorithm, the Weighted Blending algorithm has been also proposed in [10], aiming at smoothing these artifacts with a blending process. As a result, depth planes that are out of focus appear in the rendered 2D view with a more natural blurred look.

Basically, the rationale behind the Weighted Blending algorithm is the fact that neighboring MIs capture overlapping areas of the scene. Then, the blending consists in averaging together all these overlapping areas across different MIs, and weighting differently the pixels depending on their relative position (inside the MI). As illustrated in Figure 2.10c, each MI is overlapped with a shift of PS_h pixels (corresponding to the patch size) to its neighboring MIs in horizontal directions and, similarly, with a shift of PS_v in the vertical directions. Then, overlapping pixels across various MIs are averaged by using a specific weighting function. This weighting function resembles a bivariate Gaussian function (non-normalized) of size $MI_h \times MI_v$, whose mean vector, μ, is determined by the patch position, and the covariance matrix, Σ, can be used as an additional parameter to control the level of blur in out of focus areas. Hence, the weight applied to a pixel in the position $\mathbf{x} = (x, y)$ inside its corresponding MI is given by (2.3). As in the Basic Rendering algorithm, by varying the patch size and patch position parameters, it is possible to render different output views with resolution of $(PS_h \times MLA_h)\times(PS_v \times MLA_v)$.

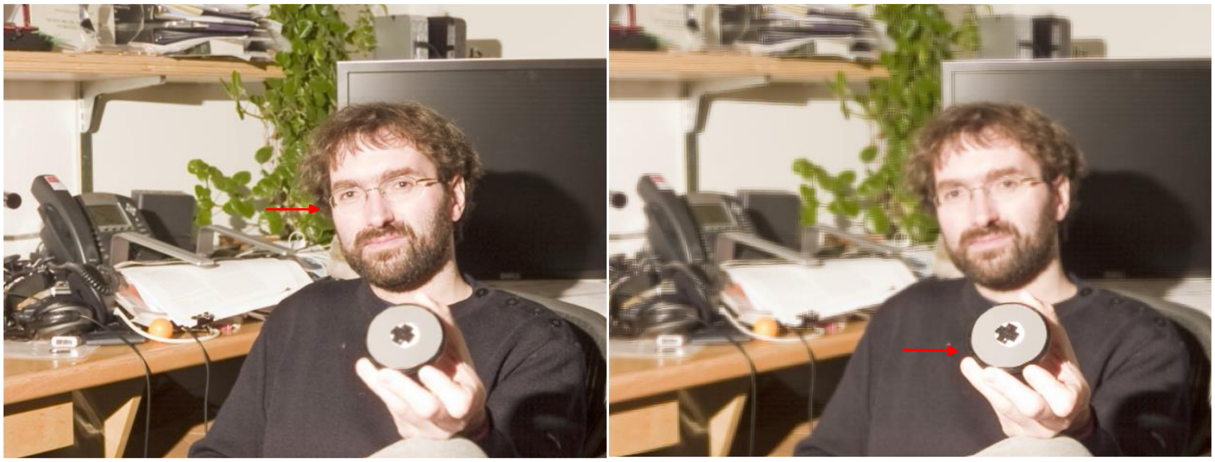
$$Weight(\mathbf{x} \mid \mu, \Sigma) = Gaussian\left(\mathbf{x} \mid \mu, \Sigma\right) = \alpha\left(\frac{1}{2}\left(\mathbf{x}-\mu\right)^{\mathrm{T}} \Sigma^{-1}\left(\mathbf{x}-\mu\right)\right) \tag{2.3}$$

Figures 2.11-2.13 illustrate rendering results when using different algorithms and/or different rendering parameters. For this, Figure 2.11 shows two 2D views rendered from the LF image *Fredo* (see Figure A.1) using the Basic Rendering and Weighted Blending algorithms with the same rendering parameters. In addition, Figure 2.12 shows two 2D views rendered from *Fredo* using the Weighted Blending algorithm with two different patch sizes. Finally, Figure 2.13 illustrates two 2D views rendered from *Fredo* using the Weighted Blending algorithm with two different patch positions.



(a) (b)

Figure 2.11 Comparison of two 2D views rendered from the LF test image Fredo (see Figure A.1) using the same rendering parameters, but different rendering algorithms: (a) Basic Rendering; and (b) Weighted Blending. The out of focus area is indicated by a red rectangle to highlight the differences between both algorithms.



(a) (b)

Figure 2.12 Comparison of two 2D views rendered from the LF test image Fredo (see Figure A.1) using the Weighted Blending algorithm with two different patch sizes to focus in the depth plane indicated by a red arrow: (a) 11×11; and (b) 15×15

2.5 LF Display

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

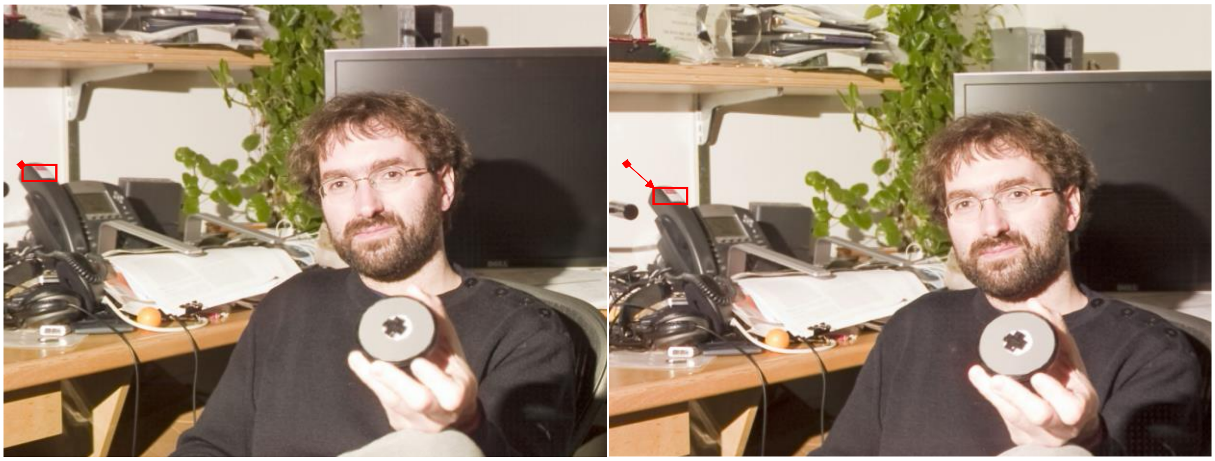
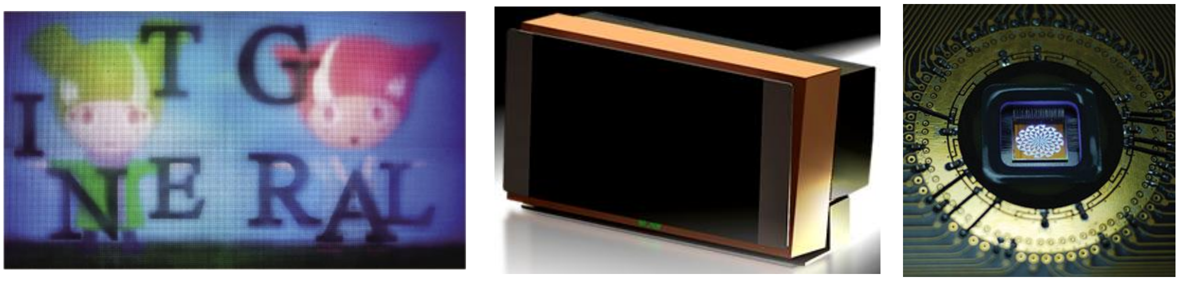


Figure 2.13 Comparison of two 2D views rendered from the LF test image Fredo (see Figure A.1) using the Weighted Blending algorithm with two different patch positions (relative to the MI center): (a) (-18, -18); and (b) (18, 18). The disparity of the captured phone in the two views is highlighted by a red arrow.

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

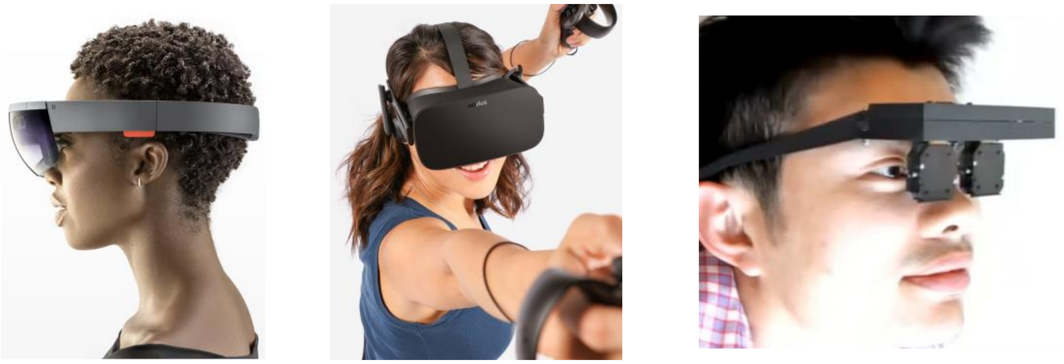
Moreover, following the recent developments in sensor and optical manufacturing, the display technologies are also evolving for providing a more natural and immersive visualization. Therefore, some prospective display technologies have also started to show up. Among them, it is possible to cite:

- **LF Displays** – A display technology using an optical setup similar to the one used for LF capturing can be also designed for LF visualization, as proposed by Nippon Hōsō Kyōkai (NHK) Japan Broadcast Corporation [14–16] (see Figure 2.14a). In this case, a flat panel display overlaid with a MLA is used. Thus, the intersection of the light rays passing through the MLA is able to re-create the full optical model of the captured scene in front of the display, which can then be observable (without eyeglasses) with continuous full motion parallax (in horizontal and vertical



(a) (b) (c)

Figure 2.14 Examples of Light Field displays: (a) NHK 8K UHD Integral Display prototype [16]*; (b) Holografika HoloVizio 80WLT HD display* [18]*; and (c) Ostendo QPI prototype* [101, 102]



(a) (b) (c)

Figure 2.15 Examples of AR and VR HMDs: (a) Microsoft HoloLens [103]*; (b) Oculus Rift VR* [6]*; and (c) NVIDIA Near-Eye Light Field HMD prototype* [24]

directions) throughout the viewing zone and with a more accurate accommodation-convergence sensation (with regard to stereo and multiview). In this case, the full LF content that was acquired needs to be delivered to the display. Another LF display technology is the so-called super-multiview LF displays, as proposed by Holografika [5, 17, 18] (see Figure 2.14b) and Ostendo [101, 102] (see Figure 2.14c), which uses a very dense number of views to create a replica of the 4D light field.

- **AR and VR Displays** – AR and VR Head Mounted Displays (HMD) allow the user to see different perspectives as he/she moves through the scene. In the case of an AR HMD (see Figure 2.15a), the real environment is seen through half-transparent mirrors and then virtual 2D views are seamlessly blended into the real scene [23, 103]. In the case of a VR HMD (see Figure 2.15b), a large number of virtual 2D views are delivered to the HMD for providing to the user the impression of immersion in a real environment [71]. Some authors have also shown that it is possible to take advantage of the microlens-based LF imaging technology for generating natural LF content for AR systems [23], as well as for creating a more natural visualization in VR HMDs [24] (see Figure 2.15c).

2.6 LF Quality Evaluation

The emergence of LF acquisition and display devices in the consumer market is relatively new and, consequently, the discussion about quality of experience for LF imaging applications is at a very early stage. Therefore, although being out of the scope of this Thesis, proposing objective and subjective quality evaluation metrics for this type of content is still an open issue, and despite its paramount importance, there have been only a few works addressing it in the literature.

In [104], an angle-dependent objective metric is proposed to be compatible with the viewing characteristics of LF content. Instead of using the general Peak Signal to Noise Ratio (PSNR) calculated over the entire LF image, the average PSNR across all its VIs is taken, as well the corresponding minimum and maximum PSNR values. In [93], two metrics are proposed for objective quality assessment in LF content: i) sparse angle-dependent; and ii) sparse depth-dependent. In both metrics the explicit knowledge of the optical capturing system (e.g., focal length, microlens pitch and MLA position) is used to extract views from the LF content at the exact position of a (simulated) viewer in front of a microlens-based LF display. Therefore, in the sparse angle-dependent metric, the average Structural Similarity Index (SSIM) is taken in a set of five views from equidistant viewing angles (with fixed distance to the MLA). In the sparse depth-dependent, depth information is estimated and an average PSNR is taken in a rendered view (with fixed angular position) over sets of pixels in different depth layers. In [105], an MI deviation-dependent objective metric is proposed. As in [93], the authors consider that the LF content is viewed using a microlens-based LF display. Therefore, based on the fact that a viewed scene is formed by the intersection of the light rays passing through neighboring microlenses, the average PSNR and standard deviation across all micro-images are calculated.

More recently, following the activities of the JPEG Pleno standardization initiative [35], as well as the ICME 2016 grand challenge on light field compression [106], methodologies for subjectively assessing the visual quality of LF content have started being addressed in the literature. In [106], a methodology is proposed for visually evaluating compressed Lytro Illum LF images. For this, views with different viewing angles and focused in different planes are rendered from decoded LF images by using the Matlab Light Field Toolbox [107]. Then, these views are visualized in a conventional 2D display using Double Stimulus Continuous Quality Scale (DSCQS) metric [108]. For this, the decoded LF images are presented simultaneously with the uncompressed reference in a side-by-side fashion, and its position on the screen is randomized. After a training session, subjects are asked to rate the quality of both images (knowing that one of them is always the reference) and the results are then analyzed using Mean Opinion Score (MOS). In [109], a methodology as well as prototype software for performing subjective quality assessment of compressed Lytro Illum LF images is also proposed. The proposed solution aims at designing a methodology that enables global assessment of quality of experience in a flexible and interactive way [109].

For this, interaction is enabled by using a Graphical User Interface (GUI) that allows rendering views in real-time according to a set of parameters (i.e., the viewing angle and the plane of focus) chosen by the user. Hence, various stimuli comparison methods [109] are adopted to obtain scores for the test material, and the results are then analyzed using MOS.

2.7 Final Remarks

This chapter presented the principles of light field imaging and reviewed the basics and recent developments related to light field imaging acquisition, representation, rendering, display, and quality evaluation. Moreover, the chapter gave a particular emphasis on the microlens-based light field imaging technology, and, in this context, reviewed some terms and concepts that are relevant for understanding the following chapters and which will be referred back to throughout this Thesis.

Chapter 3

Light Field Image and Video Coding: A Review of the Literature

Although the topic of LF content coding only recently became a more active and relevant research area (with the emergence of LF image and video coding standardization initiatives [31, 41]), various LF coding solutions have been proposed in the literature in the last two decades, following the advancement of digital multimedia systems and coding technologies. This chapter provides a comprehensive overview of the most relevant LF image and video coding solutions proposed in the literature in this period.

In this context, most of the proposed LF coding solutions made use of a hybrid coding architecture due to its effectiveness for providing high efficiency compression for both image and video. Therefore, Section 3.1 presents the basic concepts of hybrid coding, which will help the reader better understand the relevant characteristics of each LF coding approach existing in the literature. In addition to this, many of the authors adopted a standard 2D or 3D coding solution as the benchmark for experimentally evaluating their LF coding solutions. For this reason, Section 3.2 overviews the recent state-of-the art standard in 2D video coding, HEVC, while Section 3.3 briefly overviews the most relevant standards for 3D video compression. At this point, it is important to mention that HEVC and its extensions for 3D video coding are of direct relevance for this Thesis since all LF coding solutions that are here proposed make use of these standards as the basis architecture for implementing the proposed novel coding tools. Hence, it is worthwhile to emphasize that many terms and concepts that are introduced in Sections 3.2 and 3.3 will be referred back to throughout this Thesis.

After the aforementioned overview of the most relevant aspects of 2D and 3D content compression, it is then possible to better characterize the existing LF image and video coding approaches in the literature. To facilitate this, the various LF content coding approaches are clustered in the two classes depicted in Figure 3.1, by identifying which functional part of the codec is responsible for exploiting the inherent LF correlations. Notably:

- **Transform-Based Approaches** – As its name suggests, transform-based approaches exploit the LF correlations in the transform domain, based on a particular transform coding technique. LF coding solutions in this class are reviewed in Section 3.4.

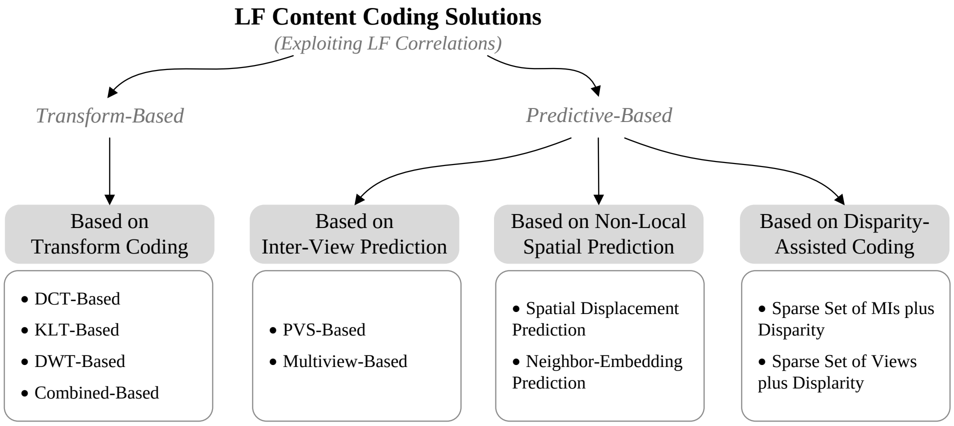

LF Content Coding Solutions
(Exploiting LF Correlations)
Transform-Based
Predictive-Based
Based on Transform Coding
Based on Inter-View Prediction
Based on Non-Local Spatial Prediction
Based on Disparity-Assisted Coding
• DCT-Based
• KLT-Based
• DWT-Based
• Combined-Based
• PVS-Based
• Multiview-Based
• Spatial Displacement Prediction
• Neighbor-Embedding Prediction
• Sparse Set of MIs plus Disparity
• Sparse Set of Views plus Displarity

Figure 3.1 The LF content coding solutions reviewed in this Thesis are grouped into four categories, according to the particular way the LF inherent correlations are exploited. Each identified category, highlighted in the shaded gray blocks, is reviewed in Sections 3.4 to 3.7 (including the sub-categories shown in the corresponding bullet points)

- **Predictive-Based Approaches** – Differently, predictive-based approaches exploit the LF correlations in a predictive manner. As illustrated in Figure 3.1, predictive-based approaches can be further categorized depending on the particular data format and prediction schemes that are adopted. Notably, three categories are identified: i) LF content coding based on inter-view prediction, which is reviewed in Section 3.5; ii) LF content coding based on non-local spatial prediction, reviewed in Section 3.6; and, finally, iii) disparity-assisted LF image coding, reviewed in Section 3.7.

It is important to highlight that, although transform- and predictive-based approaches appear separated from each other in Figure 3.1, it does not mean that a transform-based approach excludes completely any type of predictive coding tool from its architecture (or vice-versa). For instance, a predictive-based approach may use a hybrid coding architecture, as the one seen in the following section.

3.1 The Hybrid Coding Architecture

The most successful class of visual coding architectures is the one known as hybrid coder, which has been adopted in most video coding standards from H.261 [110] to the more recent H.264/AVC [38] and the state-of-the-art HEVC [2] standards.

The block diagram of a conventional hybrid encoder is illustrated in Figure 3.2. This model is called hybrid as it combines the advantages of using a transform coding stage from classical still image coding solutions (see Figure 3.2), such as JPEG [111] and JPEG 2000 [112], with a prediction modeling loop (see Figure 3.2), in which a prediction signal is generated from information available at both encoder and decoder sides.

The hybrid coding architecture has proven to be well-suited for both video and still image coding [113]. Moreover, most of the LF image and video coding frameworks presented in the literature, including the solutions proposed in this Thesis, have been based on this framework due to its effectiveness for providing high efficiency compression.

For this reason, this section reviews the fundamentals of 2D image and video hybrid coding, starting with a brief characterization of the input signal that is used to feed the encoding process, and following to the main functional blocks of the hybrid coding architecture depicted in Figure 3.2.

3.1.1 Input Signal Representation and Block Partitioning

As discussed in Chapter 2, the raw image and video may undergo multiple pre-processing operations in preparation for the encoding process. These operations include, for instance, the adaptation of the input content to a certain data representation and color format that is more suitable for the target application scenario.

Regarding the possible color formats, most capturing and display systems make use of RGB (Red, Green, and Blue) for representing color, since the HVS has three types of cones with maximum sensitivity in the wavelength areas of R, G, and B. Differently, the Y’CbCr (Luma, Blue-Difference Chroma, and Red-Difference Chroma) color format is commonly preferred for coding and transmission [114] of digital image and video. Consequently, a picture is typically converted from RGB to Y’CbCr before coding, as specified in ITU-R BT.601 [115] / BT.709 [116] / BT.2020 [117], developed by the International Telecommunications Union (ITU) – Radiocommunication Sector (ITU-R). This is mainly since Y’CbCr enables to explore the fact that the HVS system is much more sensitive to variations in luma than in chroma in order to reduce the resolution allocated to the three components. This process is known as chroma subsampling, and is usually expressed by the relation between the number of luma samples compared to the number of chroma samples. The most common are the 4:2:2 and 4:2:0 Y’CbCr subsamplings that require, respectively, two-thirds and half of the resolution of the complete 4:4:4 Y’CbCr format.

Afterwards, the encoding process operates in a block-based manner, in which the complete picture (frame) is partitioned into non-overlapping blocks and then each block feeds the encoding process (see Figure 3.2). Basically, the encoding process combines three main functional operations: i) prediction modeling, ii) transform coding, and iii) entropy coding [118] – which are all explained in the following.

3.1.2 Prediction Modeling

The prediction modeling aims at reducing the redundancy by exploiting the inherent correlations of the input video. For this, a prediction is formed by using spatially neighboring samples – known as intra prediction (see Figure 3.2) – or by using neighboring frames – known as inter prediction (see Figure 3.2). Traditionally, inter prediction was designed for

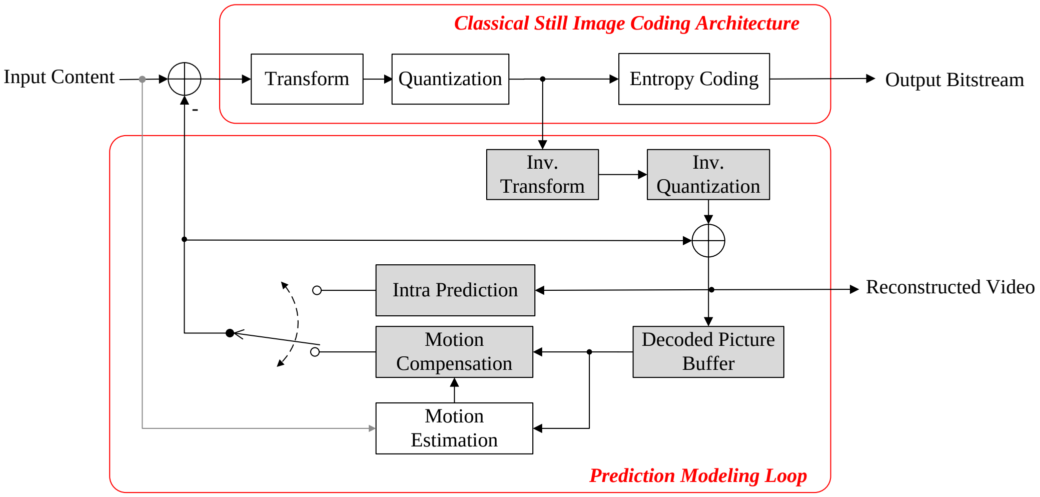


Figure 3.2 Block-diagram of a conventional hybrid video encoder (intra prediction block has been available since the MPEG-4 standard [118]*). Built-in decoder is shown in gray shaded blocks*

exploiting the redundancy between neighboring temporal frames, however, it can actually be generalized to other types of redundancy (e.g., see the inter-view prediction in 3.3.1).

As was discussed in the Chapter 2, image samples within a local neighborhood are most likely to be highly correlated. Therefore, the intra prediction block is typically built by extrapolating the pixels along the top and/or left edges of the current block (assuming that the blocks are coded in raster scan order). The output of this process is the residual block, formed by subtracting the prediction from the current block, together with an indication of how the prediction was generated.

On the other hand, in inter prediction, a motion compensated prediction (see Figure 3.2) is used for modeling the translational moving blocks in different frames. In this case, a displacement vector (known as motion vector) is used to indicate the horizontal and vertical positions (relative to the current block position) of the prediction block inside a reference picture. The motion vector may point to integer or fractional pixel positions inside the reference picture. For this, the prediction information in subsampled pixel positions are generated using interpolation filters. The reference picture is chosen from a list of frames that are stored and available in the decoded picture buffer (see Figure 3.2). Regarding the prediction block, it can be generated from a single prediction signal in a reference picture (known as uni-prediction) or from prediction signals in two reference pictures (known as bi-prediction). The process to derive a prediction block and to generate the inter prediction data can be summarized as follow:

- **Motion Estimation** – The motion estimation is used to try to estimate the prediction block with the highest similarity to the current block in a reference picture. This process is only performed at the encoder side and does not have normative restrictions. However, commonly, a block-based matching algorithm is used, in which the candidate block that minimizes a suitable matching criterion is chosen as the 'best' candidate predictor. Thus, the 'best' candidate predictor becomes the prediction for the current block. If bi-prediction is used, the two 'best' candidate predictor blocks are found (one from each reference picture) and are then linearly combined to derive the prediction block. It is worth noting that motion estimation is one of the most computational demanding operations done at the encoder side. For this reason, instead of applying the matching algorithm to the entire reference picture, a limited search window is usually considered to reduce the computational complexity.

- **Motion Compensation** – The estimated prediction block is subtracted from the current block to form the residual block. Therefore, this residual block together with motion vector(s) and some additional signaling information forms the output inter prediction data that are encoded and transmitted.

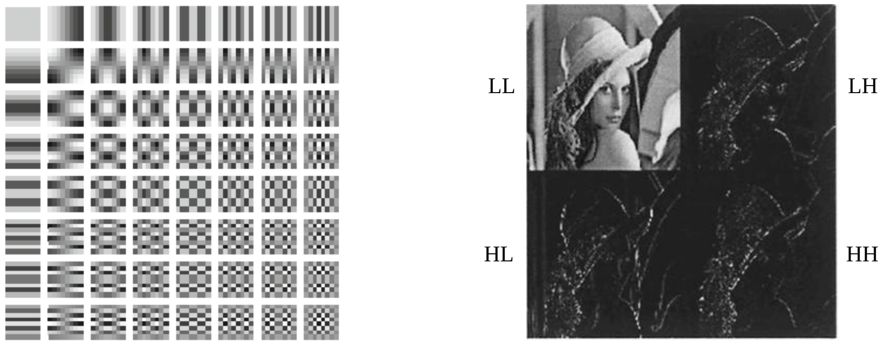
Typical hybrid video encoders achieve their compression performance by adaptively employing their multiple coding options for each region of the input picture. These coding options may comprise the applicable prediction mode (e.g., intra or inter prediction), the corresponding prediction data, and the applicable block partitioning used for prediction and transform coding. For instance, this means that an optimized encoder should prefer using intra prediction for encoding a particular region of the image in which the translational motion model breaks down.

Conceptually, the coding options are associated with signal-dependent RD characteristics, and the goal in the optimization task is to minimize the distortion, *D*, subject to a constraint on the number of used bits, *R*, over the set of all available coding options, Δ. This RDO problem is commonly solved by the Lagrangian formulation [52] in (3.1), where an optimal solution to the constrained problem is found for any value of the Lagrangian multiplier, λ (where $\lambda \geq 0$). The Lagrangian multiplier is a constant that controls the tradeoff between the quality of the reconstructed video and the bitrate of the bitstream:

$$\min_{d\in\Delta}\left\{J\right\}, \quad \text{where} \quad J(d)=D(d)+\lambda R(d)\,. \tag{3.1}$$

The Lagrangian formulation in (3.1) is commonly used by the encoder, not only for selecting the prediction mode to be used for each block, but also for solving the motion estimation process in the inter prediction. In this case, *D* may correspond to, for instance, the Sum of Absolute Differences (SAD) or the Sum of Square Differences (SSD), and *R* corresponds to the estimated number of bits for encoding the motion parameters. Conversely, on the decoder side (see shaded blocks in Figure 3.2), such decisions are directly extracted from the transmitted bitstream.

Due to continuous improvements in computational power, newer video coding standards support an increased set of coding options. In addition, more complex and efficient prediction schemes have also been investigated and adopted. Sections 3.2 and 3.3 are devoted



(a) *(b)*

Figure 3.3 Conventional transform coding: (a) 8×8 2D DCT basis functions; and (b) One level 2D DWT decomposition applied to an image (Lena)

to the relevant predictive techniques used in the state-of-the-art of video coding technology, the HEVC standard.

3.1.3 Transform Coding

The input for this process is the residual block (also called residue) from the prediction modeling. Different from the original input picture, the residual data typically present an autocorrelation function that drops off very rapidly [118]. Moreover, the more accurate the prediction process is, the less energy is contained in the residue. Therefore, the goal of transform coding (see Figure 3.2) is to convert the residue into the frequency domain such that it has a representation that is both decorrelated – i.e., separated into components with minimal inter dependence – and compact – i.e., where most of the energy is concentrated into a small number of values.

The most effective and widely used transform in hybrid video coders is the 2D Discrete Cosine Transform (DCT). However, MPEG-4 Visual [119] standard has also adopted the 2D Discrete Wavelet Transform (DWT) for coding still texture objects [120]. These two transforms are summarized as follow:

- **2D DCT** – Basically, 2D DCT is a block transform that operates in an image block of *M*×*M* samples and creates a block of *M*×*M* coefficients representing the image block in the frequency domain. It is also possible to compute the *M*×*M* coefficients by applying a separable One Dimensional (1D) DCT in the horizontal and vertical directions. Generally, the 2D DCT has low memory requirements and is well-suited to compression of block-based motion compensation residual blocks. However, a common drawback is that it tends to present blocking artefacts [118]. The DCT coefficients can be regarded as 'weights' of a set of orthogonal basis functions, as can be seen in Figure 3.3a (for an 8×8 block). These basis functions comprise a combination of horizontal and vertical cosine functions with different spatial frequencies, which go from the most significant low frequency component at the top-left position (i.e., the DC coefficient) to the less significant high frequency component of the bottom-right AC coefficient (see Figure 3.3a). At the decoder side, an inverse 2D DCT transform (see Figure 3.2) is applied, which takes the $M \times M$ DCT coefficients and reconstructs the $M \times M$ image (or residual) block back to the spatial domain. Although the usage of a floating-point DCT (such as in JPEG [111]) suffers from irreversible damage (due to finite-precision arithmetic), some coding standards (such as H.264/AVC [38] and HEVC[2]) make use of an integer DCT approximation to turn the transform coding process reversible.

- **2D DWT** – The 2D DWT is known as an image-based transform since it operates in the entire image or in large sections of the image (known as tiles), instead of small blocks. Generally, the 2D DWT presents better performance than block-based transforms (such as DCT) for still image compression, mainly by reducing the blocking artifacts at low bitrates [118]. However, it tends to present ringing artifacts and to have higher memory requirements since the whole image (or tile) is processed as a unit. The 2D DWT applies a bank of filters to the image so as to decompose it into different decomposition levels in a dyadic (pyramidal) structure. Each of these decomposition levels contains a number of frequency bands (known as subbands). By using separable transforms that can be implemented using 1D filters that are applied firstly on the rows and then on the columns, a 2D filter bank is effectively obtained. Basically, a filter pair is used to decompose a row set of samples into lowpass subband (L) and highpass subband (H) samples, which are, then, down-sampled by a factor of two. Following this, a 1D filter pair is applied along the columns of each L and H subbands to decompose them into four subbands: low frequency content in both rows and columns (LL), low frequency in columns and high frequency in rows (LH), high frequency in columns and low frequency in rows (HL), and high frequency in both rows and columns (HH). Then, each subband is again down-sampled by a factor of two. As illustrated in Figure 3.3b, the LL subband is a coarse scale approximation of the original image, while the rest of the subbands are detail information. Therefore, the 2D DWT is calculated by recursively applying this 2D filter bank to the lowest frequency subband signal (LL) obtained at each level of the decomposition. The inverse DWT at the decoder side essentially reverses the abovementioned order of operation, and then the input image is reconstructed by repeated up-sampling, filtering and addition. This operation is reversible with a correct choice of the filter bank. Examples of irreversible and reversible filter banks are, respectively, the Daubechies 9-tap/7-tap real-to-real filter [121] and the 5-tap/3-tap integer-to-integer [122] filter.

Afterwards, quantization (see Figure 3.2) is applied to the transformed coefficients. The quantizer is designed to discard insignificant values, such as near-zero coefficients, while preserving a small number of significant non-zero coefficients. In this process, the quantization step is used to regulate the range of the quantized values. The encoder duplicates the decoder process to guarantee that they both generate identical predictions for subsequent frames (see shaded gray blocks in Figure 3.2). Therefore, the quantized transform coefficients are re-scaled using the same quantization step (by the inverse quantization block in Figure 3.2), and are then inverse transformed resulting in the decoded approximation of the residual signal. Subsequently, the residue is added to the predictor, and the resulting reconstructed frame is stored in the decoded picture buffer (see Figure 3.2) to be used for the prediction of subsequent frames.

3.1.4 Entropy Coding

The small number of significant coefficients, as well the prediction parameters (e.g., quantized residue, and motion vectors), are entropy coded in order to remove statistical redundancy. For this, Context-based Arithmetic Binary Coding (CABAC) has shown to be a powerful method for providing a high degree of adaptation and redundancy reduction in the H.264/AVC standard. For this reason, the most recent video coding solution HEVC [2] has also adopted CABAC-based entropy coding.

In a nutshell, CABAC starts with a binarization process in which the entries are transformed to binary symbols (bins). For each bin, a suitable context model is then selected depending on the statistics of recently coded bins. Thus, each bin is arithmetic coded according to the selected context model.

The output of this process is the compressed bitstream, which can then be stored or transmitted.

3.2 HEVC Video Coding Standard

The continuous desire for further compression efficiency as well as for deploying video services with capabilities not previously supported – such as UHD resolution, and higher-dynamic range – has recently driven the development of a more powerful hybrid video coding solution, the HEVC standard. Developed by joint efforts of the ITU-T Video Coding Experts Group (VCEG) and ISO/IEC MPEG – named as Joint Collaborative Team on Video Coding (JCT-VC) – this recent coding solution is able to double the compression capabilities of the previous H.264/AVC standard while achieving similar quality [123]. Furthermore, HEVC was designed to be used in many different services across a very wide range of application environments, including also the support for parallel processing architectures [123].

In addition to this, although HEVC was designed to target mainly video applications, the HEVC Main Still Picture Profile [2] has also shown significant performance improvements in comparison to other standard still image codecs – such as those compliant with the JPEG and JPEG 2000 standards [124–126]. Moreover, similar conclusions have been also demonstrated for LF image coding, in [127, 128], where HEVC presented significantly better performance compared to JPEG and JPEG 2000.

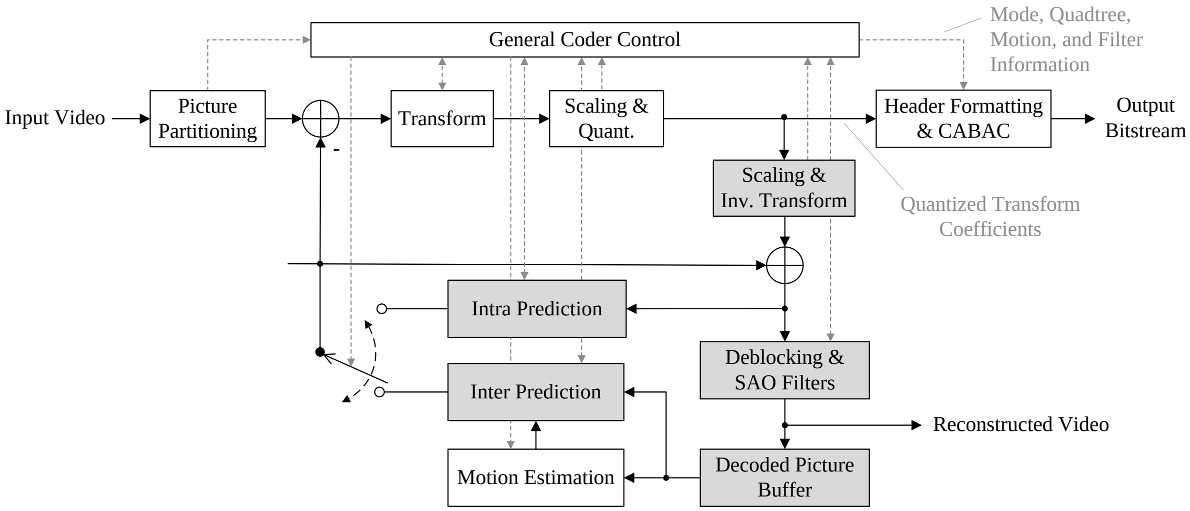


Figure 3.4 HEVC coding architecture [123] *(the built-in decoder is in gray shaded blocks)*

HEVC inherits some of the basic high-level features of H.264/AVC, such as the organization of the bitstream in Network Abstraction Layer (NAL) units and the adoption of parameter sets. The parameter sets carry high-level control information regarding the entire video sequence or regarding a set of one or more video frames. Moreover, as can be seen in Figure 3.4, which illustrates a typical coding architecture for producing an HEVC compliant bitstream, the video coding layer employs the same hybrid approach as explained in the previous section (see Figure 3.2).

HEVC is of direct relevance for this Thesis since all image and video solutions that are here proposed consider a coding architecture based on this standard. For this reason, the HEVC coding tools that contributed to the proven high efficiency of video compression are reviewed in the following sections. Notice that, this review only considers the blocks in Figure 3.4 that have normative restrictions (i.e., motion estimation is not included). For a more complete description of HEVC, please also refer to [123].

3.2.1 Picture Partitioning

A hierarchical block partitioning concept is used in HEVC to customize the partitioning for better dealing with High Definition (HD) and UHD video, as well as for adapting the block partitioning to the local properties of the picture. This very flexible partitioning is based on the following elements [123]:

- **Coding Tree Block (CTB) and Coding Tree Unit (CTU)** – Each picture is partitioned into square-shaped CTBs such that the resulting number of CTBs is identical for both luma and chroma components. The CTB (of luma samples) can be as large as 64×64 samples. Therefore, each CTB of luma together with the CTBs of chroma samples and syntax associated with these samples forms a CTU. The CTU represents the basic processing unit in HEVC.

- **Coding Block (CB) and Coding Unit (CU)** – Following a quadtree syntax (also known as coding tree), a CTU can be split into multiple CUs of variable sizes. Similar to the CTU, a CU consists of a square block of luma samples, the two corresponding blocks of chroma samples, and the associated syntax. These luma and chroma blocks are referred to as CBs. The minimum CU size can range down to 8×8 blocks of luma samples. The CUs inside a CTU are coded in a depth-first order (also known as z-scan order) so as to ensure that, for each CU (except the first), all samples above and left of it have already been coded.
- **Prediction Block (PB) and Prediction Unit (PU)** – The decision whether to use intra prediction or inter prediction is made at the CU level. Hence, a prediction mode is signaled in the bitstream for each CU to indicate the choice. For CUs that have the minimum CU size and are coded using intra prediction, the luma CB can also be decomposed into four equally-sized CBs, in which a different intra prediction mode can be signaled for each CB. If a CU is coded using inter prediction, the luma and chroma CBs can be further split into one or more PBs. In this case, each luma and chroma PB is a block that uses the same motion parameters for motion compensated prediction and the same PB splitting is used for all color components. The luma PB, together with the chroma PBs and associated syntax form a PU. For each PU, a single set of motion parameters is signaled in the bitstream. HEVC supports eight different modes for partitioning a CU into PUs, as illustrated in Figure 3.5 (assuming a CB of $M \times M$ luma samples).
- **Transform Block (TB) and Transform Unit (TU)** – The block-based transform coding is also performed using a quadtree partition approach. In this case, each CB can be recursively divided into multiple TBs. A TB represents a square block of color component samples in which the same 2D transform is applied for coding the residual data. The luma TB size can go from 4×4 up to 32×32 samples. A TU represents a luma TB greater than 4×4 samples (or four luma TBs with 4×4 samples each), the corresponding two chroma TBs, and associated syntax. This additional degree of freedom on deciding how to partition into TBs allows adapting the transform basis function to the varying space-frequency characteristics of the residual signal.

The decision of how to partition the CTB into the aforementioned block elements is made using the Lagrangian RDO approach, similarly to the one mentioned above in Section 3.1.2.

In addition to the block partition, the high-level segmentation of the picture in HEVC is achieved by using the concept of slices and slice segments. A slice can consist of the complete picture or parts thereof. Also, each slice is formed by one or more slice segments. A slice segment comprises a set of CTUs that can be independently or dependently decodable from CTUs in other slice segments [123]. Similar to the concept in H.264/AVC, a slice segment can be classified as:

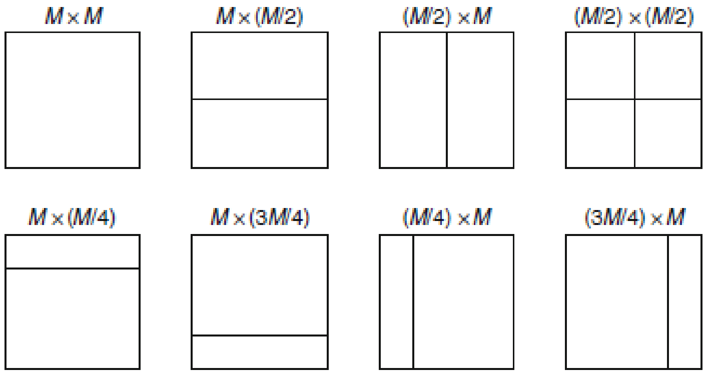


Figure 3.5 Supported partitioning modes for splitting a coding unit (CU) into one or more prediction units (PU) [123]. *For decreasing the computational time for testing all modes, the (M/2)×(M/2) and the PUs in the bottom row may be not supported for all CU sizes*

- **I Slice** – A slice segment in which all CUs are coded using only intra prediction.
- **P Slice** – A slice segment in which the CUs can be coded using either intra prediction or inter prediction with at most one motion compensated prediction signal per PB (i.e., uni-prediction). HEVC uses two separate reference picture lists to store reference frames, namely, List 0 and List 1. However, P slices only consider reference frames that are in List 0 for prediction.
- **B Slice** – A slice that, in addition to the modes available in P slices, allows the use of inter prediction with bi-prediction. In this case, each of the two reference frames that can be used in bi-prediction is separately stored in List 0 and List 1.

HEVC has also included two new tools aiming at facilitating high-level parallel processing [123], namely, Wavefront Parallel Processing (WPP) and tiles. Both of these tools allow subdivision of each picture into multiple partitions that can be processed in parallel. Each partition contains an integer number of CTUs that may (in the WPP) or may not (in the tiles) have dependencies on CTUs of other partitions. These tools may be used in both encoder and decoder sides, and their usage is signaled in the high-level parameter sets.

3.2.2 Intra Prediction

HEVC intra sample prediction is performed by extrapolating sample values from adjacent reconstructed blocks at left and top positions, as illustrated in Figure 3.6. HEVC allows using the complete set of intra prediction modes regardless of the availability of these neighboring reference samples. If there are unavailable reference samples (e.g., the samples are outside the picture, slice or tile), HEVC simply replaces these values with the closest available reference sample value.

The set of defined prediction modes consists of methods for modeling various types of content typically present in video and still images. These are:

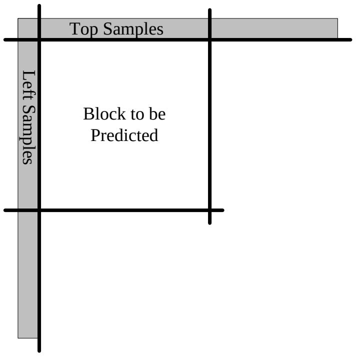


Figure 3.6 Example of reference samples to be used by HEVC Intra Prediction

- **Angular Prediction Modes** – The angular prediction is designed to provide high-fidelity predictors for content with directional structures. For this, a set of 33 angular prediction directions at 1/32 sample accuracy has been selected to provide a good trade-off between encoding complexity and coding efficiency in typical image and video content.
- **DC Prediction Mode** – The DC prediction is available for modeling smooth areas of the picture. In this case, the predicted samples are given as the average of left and above samples of the block to be predicted. Additionally, DC predicted luma blocks with size equal to or smaller than 16×16 go through a post-filtering process to soften the discontinuities that can occur on both top and left boundaries of the block.
- **Planar Prediction Mode** – The planar prediction can also be used as an alternative mode to overcome the blockiness that can still be observed in smooth areas when DC prediction is applied. This is achieved by averaging a horizontal and vertical linear predictions.

Moreover, an adaptive filtering process (referred to as low-pass smoothing in [114]) can also be used to pre-filter the reference samples according to the intra prediction mode, block size and directionality. This smoothing process intends to improve visual appearance of the prediction block by avoiding steps in the values of reference samples that could potentially generate unwanted directional edges in the prediction block.

Due to the very large amount of different block sizes and prediction mode combinations, all the intra prediction modes have been designed in such a way as to allow easy algorithmic implementations for arbitrary block sizes and prediction directions.

3.2.3 Inter Prediction

Similarly to the previous video coding standards, block-based motion compensated prediction is used to explore temporal correlations between pictures. Figure 3.7 shows the

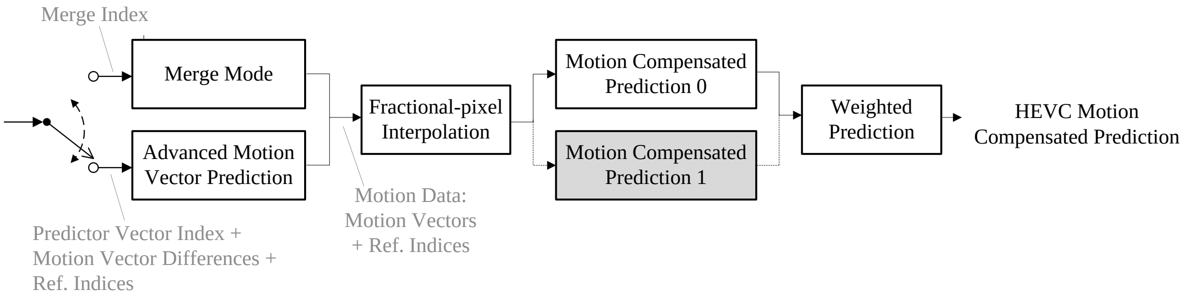

Merge Index
Merge Mode
Advanced Motion Vector Prediction
Predictor Vector Index + Motion Vector Differences + Ref. Indices
Fractional-pixel Interpolation
Motion Data: Motion Vectors + Ref. Indices
Motion Compensated Prediction 0
Motion Compensated Prediction 1
Weighted Prediction
HEVC Motion Compensated Prediction

Figure 3.7 HEVC Inter prediction [123] *(block used only for bi-prediction is in shaded gray)*

basic HEVC inter prediction walkthrough for each CU. This diagram illustrates in more detail the inter prediction block from Figure 3.4.

Basically, motion data may be derived at the encoder using the motion estimation process. In this case, in addition to the motion vectors, HEVC also transmits indices with the position of the used reference frames in the reference picture lists. These inter prediction data (among some further prediction data that will be disclosed in this section) are the input to the inter prediction process illustrated in Figure 3.7. Since motion estimation is not specified within the HEVC standard, it is bypassed in Figure 3.7 and will not be further mentioned in this section.

Besides using a much more flexible partition scheme (see Section 3.2.1), HEVC has also introduced several improvements in the inter coding tools with respect to the previous H.264/AVC standard. Notably, both H.264/AVC and HEVC use motion vectors with fractional pixel accuracy to more accurately capture continuous motion (up to quarter-pixel for luma component). However, to generate these fractional pixels, HEVC uses interpolation filter-kernels with higher precision than H.264/AVC, i.e., seven/eight-tap filters for luma and four-tap filter kernel for chroma. These larger tap-filters have improved the filtering process, especially in the high frequency range [123].

Moreover, the weighted prediction signaling was also simplified in HEVC by explicitly signaling whether different weights are applied to each motion compensated prediction or the two motion compensated predictions are just averaged with equal weights (i.e., equal to $1/2$).

In addition to this, since motion vectors of a specific block are likely to be correlated with the motion vectors in neighboring blocks, they are not directly coded in the bitstream, but predictively coded based on their neighboring motion vectors. This predictive coding of the motion vectors was improved in HEVC by introducing a new tool called Advanced Motion Vector Prediction (AMVP). Furthermore, a new HEVC technique called merge mode is used to derive all motion data of a block (i.e., motion vectors and indices of the used reference pictures) from neighboring blocks, replacing the direct and skip modes of H.264/AVC. These two new techniques are discussed in more detail in the following.

3.2.3.1 AMVP

Each motion vector is coded as a difference (i.e., the motion vector difference shown in Figure 3.7) with respect to a motion vector predictor. This motion vector predictor is derived from already decoded motion vectors from spatial neighboring blocks or from temporally neighboring blocks (known as co-located blocks).

AMVP was designed to derive the motion vector prediction and then to signal it explicitly to the decoder side. For this, a list of candidate motion vector predictors is constructed and the index to the chosen candidate in the list is included into the bitstream (i.e., the predictor vector index shown in Figure 3.7). During the HEVC standardization process, many simplifications were investigated to reduce the HEVC computational complexity and signaling without sacrificing the coding efficiency. These simplifications led to the following AMVP candidate list [123]:

- **Spatial Candidates** – Up to two spatial candidates are derived from a set of five spatial neighboring blocks, namely: left-bottom (A_0), left (A_1), above-right (B_0), above (B_1), and above-left (B_2), as illustrated in Figure 3.8. The first spatial candidate is derived by taking the first available motion vector found in the blocks A_0 or A_1 in this order (i.e., the motion vector of the first block, among A_0 or A_1, which was coded using inter prediction). As for the second spatial candidate, it is given by the first available motion vector in the neighboring blocks B_0, B_1, and B_2 (in this order).
- **Temporal Candidates** – When neither of the spatial candidates are available (or they are identical), one temporal candidate is derived as the first available motion vector from the two temporal co-located blocks shown in Figure 3.8. These are: bottom-right (C0), and center (C1) blocks.
- **Zero Motion Vectors** – When spatial, temporal or either of these candidates are unavailable, zero motion vectors are added to fill the candidate list with up to two final candidates.

3.2.3.2 Merge Mode

The merge mode is used to extend the concept of the direct modes from H.264/AVC standard. The merge mode allows coding the motion data (i.e., motion vectors and indices to the reference picture lists) with very low bitrate.

For this, similarly to AMVP, a list of motion data candidates is derived by using neighboring blocks. Therefore, only an index (i.e., the merge index shown in Figure 3.7) is signaled which identifies the used candidate in the list. However, the AMVP list contains motion vectors from only one reference list, while the merge list comprises the complete motion information for the two reference picture lists.

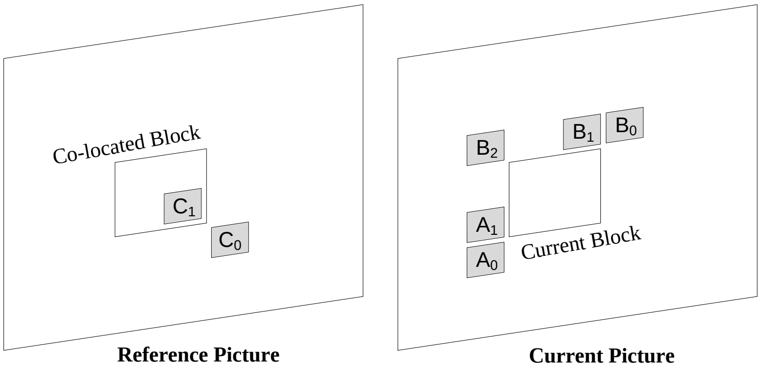


Figure 3.8 Candidates for AMVP and merge mode: spatial (A_0, A_1, B_0, B_1, and B_2) and temporal (C_0, and C_1) candidates

Briefly, the merge list is constructed with the following candidates:

- **Spatial Merge Candidates** – Up to four spatial candidates are derived from the set of five spatial neighboring blocks shown in Figure 3.8. These four candidates are taken by sequentially checking A_1, B_1, B_0, A_0, and B_2, respectively. A redundancy check is also performed to avoid candidates with equal motion information.
- **Temporal Merge Candidates** – One temporal candidate is derived as the first available motion vector from the two temporal co-located blocks shown in Figure 3.8 (C_0, and C_1, in this order).
- **Additional Merge Candidates** – The maximum size of the merge candidate list is signaled in the slice header syntax (being equal to five by default). If the merge list is not fully populated, additional candidates can be appended to the list. For B slices, bi-predicted candidates are generated by combining existing candidates from each reference picture list. When the list is still not full (or in a P slice), zero motion candidates are included to complete the list.

A crucial application of the merge mode is the generalization of the so-called skip mode from H.264/AVC standard. In the skip mode, only a skip flag is signaled, and no further information is sent throughout the bitstream (i.e., the residue is zero). In HEVC, at the beginning of each CU in a P or B slice, a skip mode can be signaled, which implies that:

1) The CU only contains one PU (no further PB partitioning).
2) The merge mode is used to derive the motion data (motion vectors and indices to the reference picture lists).
3) No residual transform coefficients are present in the bitstream.

3.2.4 Transform, Scaling and Quantization

Following the hybrid coding architecture shown in Figure 3.4, HEVC also uses transform and quantization methods to encode the residual signal.

Regarding the transform, 2D integer approximations to the DCT transform are defined with various sizes from 4×4 up to 32×32. In addition, HEVC also specifies an alternative 4×4 integer transform, based on the Discrete Sine Transform (DST), to be used with 4×4 luma intra prediction residual blocks. These DST basis functions were shown to perform better (in terms of bitrate reduction) than the DCT basis functions in modeling the spatial characteristic of the intra prediction residual [129].

Concerning the quantization, HEVC uses a quantizer design similar to H.264/AVC. In this case, a Quantization Parameter (QP) in the range of 0 to 51 is mapped to a quantizer step size that doubles each time the QP value increases by 6 [118]. In addition to this, in HEVC, a QP value can be transmitted (in the form of delta QP) for a quantization group as small as 8×8 samples for rate control and perceptual quantization purposes. The QP predictor used for calculating the delta QP uses a combination of left, above and previous QP values.

3.2.5 Deblocking and Sample Adaptive Offset Filters

A filtering process is applied to the reconstructed samples before writing them into the decoded picture buffer shown in Figure 3.4. This process is applied in both encoding and decoding loops. The filtered picture is then used for motion compensation of subsequent pictures, and generally improves the compression performance, being a more faithful reproduction of the original image (compared to the unfiltered picture). HEVC specifies two in-loop filters, notably, a deblocking filter and a Sample Adaptive Offset (SAO).

The deblocking filter aims at reducing the blocking artifacts by attenuating discontinuities at the PBs and TBs borders. Discontinuities are mainly caused by motion prediction of adjacent blocks that may come from non-adjacent areas of the reference picture. In addition to this, coarse quantization can also create discontinuities at the block boundaries.

On the other hand, the SAO is a non-linear filter that is applied to the output of the deblocking filter and aims at attenuating the ringing artifacts as well as possible changes in sample intensity in some areas of the picture. Ringing artifacts are more likely to appear when larger transform sizes are used. Briefly, SAO firstly classifies the samples in a region into multiple categories and adds a specific offset to each sample, depending on its category. Both the classifier index and the used offset are signaled in the bitstream (for each CTB). HEVC uses two types of SAO classifiers, known as edge-offset and band-offset. The edge-offset compares the current samples with its neighboring samples and classifies according to the gradient patterns. In the band-offset, the classification is based on the samples amplitude values.

3.2.6 Decoded Picture Buffer

Pictures in the decoded picture buffer can be marked as “used for short-term reference”, “used for long-term reference”, or “unused for reference”. This process of marking pictures is managed using the Reference Picture Set (RPS), which is signaled for each slice in the slice header syntax [123].

Basically, the RPS comprises Picture Order Count (POC) values of the pictures that must be kept in the decoded picture buffer to be used as a reference to current or future pictures. Hence, pictures that are indicated in the short-term or long-term portion of the RPS are kept, respectively, as short term and long-term pictures. The difference between both is that a long-term picture can be kept in the decoded picture buffer much longer than a short-term picture [123]. On the other hand, pictures in the decoded picture buffer that are not indicated in the RPS are marked as “unused for reference” (also known as non-reference). Non-reference pictures are not used as reference and are only kept in the decoded picture buffer only if they need to be output later.

3.2.7 Header Formatting and Entropy Coding

In the last stage of the HEVC encoding process, the video signal is reduced to an ordered stream of syntax elements. Then, syntax elements at the highest-level, such as NAL unit headers, are coded using fixed length coding. Following this, syntax elements at the level of parameter sets are coded using either fixed length codes or variable length codes (notably, the zero-order Exponential-Golomb coding). Differently, syntax elements at the coding block level (e.g., describing the used method of prediction, prediction parameters and residual data) are entropy coded using CABAC. However, it is worthwhile to notice that the entropy coding block illustrated in Figure 3.4 is used to generally address the operation of transforming the entire stream of syntax elements (not only the ones at the coding block level) into the output bitstream.

Concerning the HEVC implementation of CABAC, some improvements with respect to H.264/AVC have been introduced to increase the CABAC throughput and compression performance, as well as to reduce its context memory requirements. Details about design considerations for CABAC in HEVC can be found in [123].

3.3 3D Video Coding Standards

As previously explained in Chapter 2 (see Section 2.3), an LF image can be reorganized as a set of MIs, or VIs that can be used to adopt a multiview representation. Moreover, depth/disparity information can also be extracted from the LF content, which can be further represented by using a multiview plus depth (or disparity) format. Some of the LF coding schemes in the literature have adopted these types of approaches, as will be further discussed in Sections 3.5 and 3.7. For this reason, this section briefly reviews the relevant concepts behind state-of-the-art 3D video coding standards.

3.3.1 Multiview Video Coding

The desire for an efficient representation of multiview video has originated from the recent growing interest in 3D video application services, such as 3D Blu-Ray discs. To accomplish this, both H.264/AVC and HEVC standards comprise a simple, but efficient solution for multiview video coding, respectively, the MVC [39] and the Multiview (MV)-HEVC [40] extensions.

The basic idea in these multiview video coding solutions is to exploit not only the redundancies that exist temporally between the frames within a given view, but also the similarities between frames of neighboring views (known as inter-view prediction). Notably, since the captured views in a multiview scenario typically record the same scene from nearby viewpoints, substantial inter-view redundancy is present [39]. This inter-view redundancy can then be exploited by compensating the displacement (disparity) between object positions in different views. This process is known as disparity compensation, which was earlier adopted in H.262/MPEG-2 [36] Multiview Profile [37] as shown in Figure 3.9. Notice that, the coding principles used in this codec architecture (see Figure 3.9) are mostly the same as those currently used in more advanced 3D video coding solutions [130]. By using inter-view prediction, a reduction in bitrate (24 % on average with up to 8 views [130]) relative to independent coding of these views can be achieved without sacrificing the reconstructed video quality.

In addition to this, an important requirement for adoption of a multiview coding standard by the industry was to modify as few aspects of the underlying 2D video coding standard design as possible. For this, these solutions make use of the inter prediction method also for inter-view prediction by making the decoded pictures from other views also available in the reference picture lists. Hence, similar inter prediction techniques can be used for both disparity and motion compensated prediction. Regarding the bitstream syntax, both MVC and MV-HEVC extensions are supported by making changes only in the high-level syntax, for instance, for signaling the inter-view prediction dependency. This allows reusing most of the existing implementations without major changes for building the MVC and MV-HEVC decoders.

In order to take advantage of the state-of-the-art compression capabilities achieved by HEVC, ISO/IEC MPEG and ITU-T VCEG standardization bodies have established the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) for developing next generation 3D coding standards. As part of this, MV-HEVC was integrated in the second edition of the HEVC standard [40].

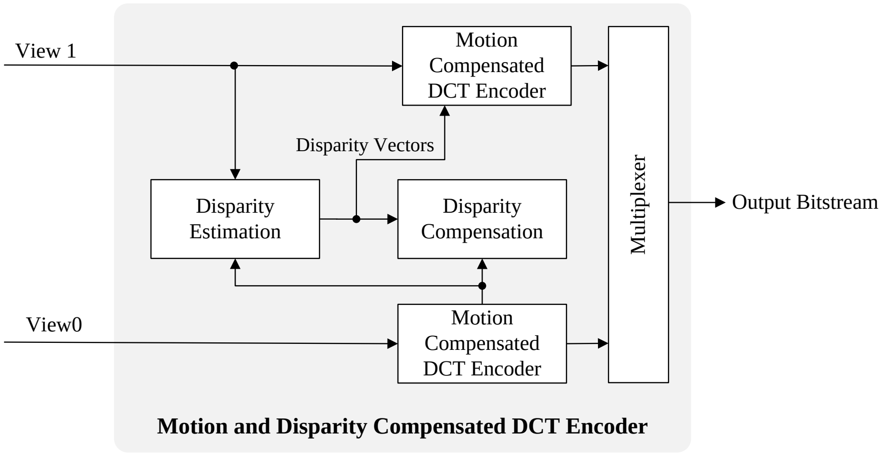

View 1
Motion Compensated DCT Encoder
Disparity Vectors
Disparity Estimation
Disparity Compensation
Multiplexer
Output Bitstream
View0
Motion Compensated DCT Encoder
Motion and Disparity Compensated DCT Encoder

Figure 3.9 MPEG-2 multiview profile coding architecture [37]

MV-HEVC employs a multilayer approach where different HEVC-coded representations of the video sequence, called layers, are multiplexed into one bitstream and can be hierarchically dependent on each other. In MV-HEVC, a layer simply represents the texture data belonging to the same camera (i.e., a view). All pictures associated with the same capturing or display time instant are contained in one access unit and have the same POC. Thus, the first picture within an access unit is usually denoted as the base layer. The pictures that follow the base layer are denoted as enhancement layers.

MV-HEVC allows having an external base layer, which is an approach that supports the base layer being coded by other codecs, e.g., H.264/AVC. Basically, except for output information and representation format, no other information about the base layer is included in the bitstream. In enhancement layer coding, MV-HEVC allows inter-layer prediction between pictures in the same access unit but in different views (i.e., inter-view prediction).

Although multiview video coding solutions provide higher compression efficiency than the independent coding of a small set of views, it has been shown that the required bitrate for providing a given level of video quality increases linearly with the number of coded views [131]. For this reason, more suitable coding solutions need to be designed for supporting 3D applications in which a larger number of views are required. An example of this are the depth-assisted 3D video coding solutions, which are reviewed in the following section.

3.3.2 Depth-Assisted 3D Video Coding

Depth-based 3D formats and coding standards allow supporting depth perception of a visual scene through a larger range of display systems, such as autostereoscopic displays, as well as allowing the viewer to freely navigate through different viewpoints of the scene [132].

Moreover, by using a multiview plus depth format, it is also possible to drastically reduce the bitrate required for the 3D video signal since only a limited number of texture views need to be transmitted, together with the corresponding depth. At the decoder side, additional intermediate views can be synthesized by using Depth Image Based Rendering (DIBR) techniques. In order to render these views with acceptable quality, it is desirable to use high-quality depth maps, which need to be represented and efficiently coded along with the texture. Depth maps can be either estimated from a multi-camera (or stereo) setup [133] or acquired by a special depth camera [134].

More recently, JCT-3V has also finalized the state-of-the-art 3D-HEVC extension, whose goal was to develop a 3D video format that could facilitate the generation of intermediate views with high compression capabilities in order to support advanced stereoscopic display functionality and emerging autostereoscopic displays [132]. For this, new coding tools were introduced, notably:

- **New Inter-Layer Coding Tools** – For further improving the MV-HEVC inter-layer coding, new types of inter-layer prediction were introduced to exploit dependencies between multiple views and between video texture and depth. Therefore, three new types of prediction are allowed: i) a combination of temporal and inter-view prediction; ii) inter-component prediction, which refers to the combination of texture and depth components that occur at the same time instant and correspond to the same view; and iii) a combination of inter-component and inter-view prediction, which refers to texture and depth components at the same time instant, but from different views. Moreover, residual, disparity, and partitioning information can also be predictively coded in an enhancement layer.

- **New Depth Coding Tools** – New depth coding tools were introduced so as to explicitly address the unique characteristics of depth maps. Notably, to better represent depth map data and to preserve sharp edges, new intra and inter coding tools were added. Regarding intra coding tools, a depth modeling mode is used to partition the depth block into non-rectangular regions and, then, to preserve the discontinuities in the depth map. In addition, a segment-wise DC is used in alternative to the transform residual coding, which can be employed only for $M \times M$ CU partitions (see Figure 3.5). Moreover, a single depth intra mode is also available to efficiently represent smooth areas of the depth map by approximating the depth block to a single depth value. Regarding inter coding tools, a motion parameter inheritance tool is used to predict the motion information for the depth map based on the motion characteristics of the corresponding texture component (in the same view).

- **View Synthesis Optimization** – The coding efficiency for depth coding is further improved by using a modified Lagrangian cost formulation for coding decisions, referred to as view synthesis optimization. Since the depth map is not directly visualized, but is rather used for view synthesis at the decoder side, a novel distortion metric is used, so as to account for the errors that the encoded depth map produces in the synthesized view.

3.4 Transform-Based LF Content Coding

After reviewing the basic concepts of image and video compression, as well as the techniques used in state-of-the-art 2D and 3D video coding standards in the previous sections, it is now possible to better characterize the LF image and video coding approaches existing in the literature according to the diagram shown in Figure 3.1.

Starting with transform-based approaches (see Figure 3.1), these correspond to LF content coding solutions that rely on transform coding for exploiting the inherent LF correlations. Specifically, various transform coding techniques can be used to decorrelate the LF image and then remove the redundant information between neighboring MIs or VIs (as discussed in Chapter 2). Therefore, it is possible to group the transform-based approaches into four categories, depending on the type of used transform (see Figure 3.1): i) DCT-based; ii) DWT-based; iii) Karhunen-Loève Transform (KLT) -based; and iv) combined-transform approaches.

This section reviews the relevant LF image coding solutions in each of the identified groups and describes a few solutions that propose to extend the transform-based methods for LF video coding.

3.4.1 DCT-Based Coding

For the earliest LF coding schemes proposed in the literature, the most natural choice has been to use the classical image coding architecture shown in Figure 3.2, and to apply an approach that resembles the JPEG standard but is based on a 3D version of the DCT transform (known as 3D DCT). For this, the geometry of the camera needs to be known so as to decompose the LF image into stacks of MIs or VIs (along the third dimension) in order to apply the 3D transform into a block volume. Therefore, not only the existing spatial redundancy within each MI/VI is exploited, but also the redundancy between neighboring MIs/VIs.

In [135, 136], the LF image is organized into stacks of 8 adjacent MIs and the 3D DCT is applied to each 8×8×8 block. Then, the resulting DCT coefficients are uniformly quantized and both DC and AC quantized coefficients are equally entropy coded by using a combination of run-length and Huffman coding. It is shown that the proposed solution [135, 136] presents significant improvements compared to JPEG for gray-level LF images. In [137], an alternative quantization strategy is proposed to replace the uniform quantizer from the previous solution proposed in [135, 136]. In this case, further improvements in compression performance are reported by optimizing the quantization matrix depending on the particular content of the image.

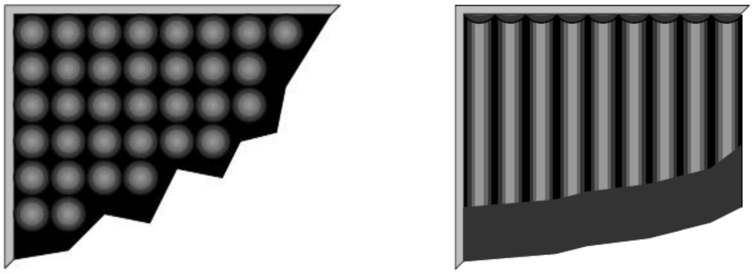


Figure 3.10 Comparison between: (a) A rectangular array with square microlenses; and (b) A lenticular-based array with cylindrical microlenses. From: Forman, 2000 [138]

It is worthwhile to notice that these coding solutions proposed in [135–137] are optimized for a special case of the LF imaging system, called lenticular-based imaging [138], where a 1D cylindrical MLA is used for capturing instead of the 2D array of microlenses, as depicted in Figure 3.10. As a consequence of this design, the resulting images contain parallax only in the horizontal direction. In [139], the solution with the 3D DCT from [135, 136] is generalized for LF with full parallax. In this case, both the horizontal and vertical MIs are decorrelated simultaneously by the 3D DCT. Hence, it is shown that different scan ordering approaches for gathering the horizontal and vertical MIs (to form 8×8×8 stacks) result in different RD performance. This fact has motivated the work in [140], which proposes to use a Hilbert space-filling curve for scanning the MIs in the array and forming 8×8×8 stacks to be 3D DCT coded (see Figure 3.11f). Various scanning topologies are compared, namely: raster, perpendicular and spiral (see Figure 3.12) and it is shown that the 3D DCT in conjunction with the Hilbert scan outperforms all other tested solutions.

In addition, an adaptive 3D DCT-based framework is proposed in [141, 142], in which the number of MIs involved in a single 3D DCT is varied (between 8×8×1, 8×8×2 , 8×8×4, and 8×8×8 stacks of MIs) according to the MI cross-correlation in a neighborhood. In this case, the MI mean values are used as a measure of correlation between MIs. Consequently, it is shown that the adaptive 3D DCT could significantly outperform the non-adaptive solution from [135, 136] (for lenticular-based images).

More recently, an alternative adaptive 3D DCT-based approach has been proposed in [143]. Similarly to [141, 142], the mean value is used as a correlation metric. However, the 3D DCT is applied to stacks of VIs and the size of the 3D DCT is varied between 16×16×16 down to 4×4×4, depending on the correlation between neighboring VIs. From this, it is shown that further improvements can be achieved compared to the adaptive solution proposed in [141, 142] (when also applied to VIs).

3.4.2 KLT-Based Coding

Instead of using the DCT as in the previous section, other schemes propose to use a KLT-based approach for LF image coding. The KLT, also known as Principal Components

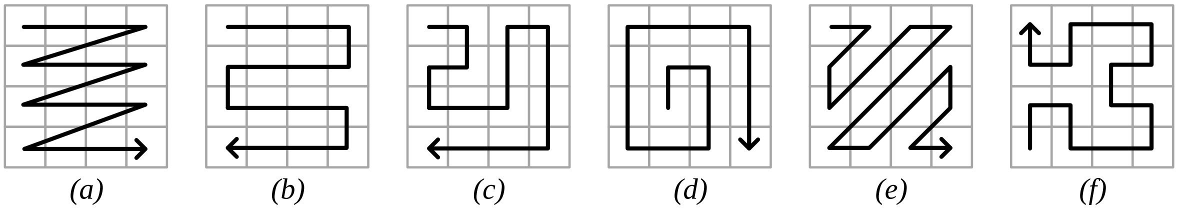


Figure 3.11 Possible scan order for arranging MIs, or VIs for coding: (a) Raster; (b) Parallel; (c) Perpendicular; (d) Spiral;(e) Zig-zag; and (f) Hilbert

Analysis (PCA) [144] decomposition and Hotelling [145] transform, is a block-based transform that exploits the statistical characteristics of the input data. The KLT consists of decomposing the input data in a set of orthonormal basis functions (known as the principal components) into which the variance of the input data is maximal. This corresponds to ordering the eigenvectors of the covariance matrix (the KLT matrix), which is calculated with the input data and ordered according to the largest eigenvalues.

The idea of applying the KLT transform for compression comes from the fact that a linear combination of any reduced number, k, of eigenvectors corresponds to the best approximation of the input data in a reduced k-dimension subspace (i.e., the approximation with minimal mean square error) [146]. Therefore, different compression ratios can be achieved by simply discarding less (or more) eigenvectors (i.e., discarding the less significant rows from the covariance matrix). Concerning the usage of KLT for image compression, although the KLT is very efficient in compacting the energy in a small number of eigenvectors, there are still some implementation related difficulties, mainly due to the fact that the KLT basis functions are image dependent. However, it may be suitable in applications where the statistics of the data change slowly and the covariance matrix is kept small [146].

Regarding LF image compression, a KLT-based coding scheme is proposed in [147], in which a Vector Quantization (VQ) scheme is used for clustering different MIs into a representative set of vectors to be then coded with KLT, as illustrated in Figure 3.12a. For this, the LF image is divided into consecutive blocks of $d \times d$ samples which are treated as a $(d \times d)$-dimensional vector. These vectors are then grouped into S different classes by using the Linde-Buzo-Gray (LBG) optimization algorithm [148]. As a result, a codebook is derived, consisting of S representative vectors (known as code-vectors). Then, a KLT with $(k \times k)$-dimension is applied into the vectors from each of the S classes so as to reduce the dimensionality of their vectors from $(d \times d)$ to $(k \times k)$-dimension vectors, where $k \leq d$. Afterwards, the reduced KLT coefficients, together with the codebook and the KLT matrix are scaled and rounded to the nearest integer to compose the output bitstream. From the presented results, it is shown that varying the $d \times d$ block size does not affect the RD efficiency for LF image coding. However, the larger is the number of sets S, the better is the observed RD performance. Moreover, the presented KLT scheme always outperforms the JPEG standard for lower bitrates.

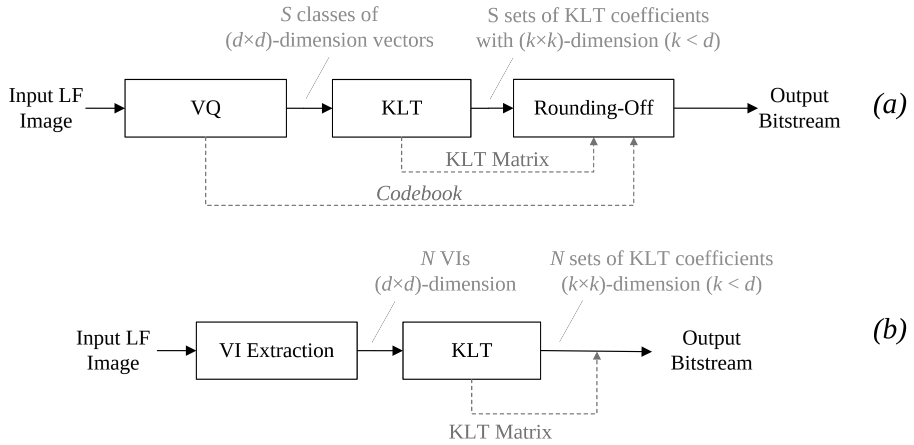


Figure 3.12 KLT-based LF image coding schemes: (a) Proposed by Jang, Yeom, and Javidi [147]*; and (b) Proposed by Kang, Shin, and Kim* [149, 150]

An alternative KLT-based coding scheme is also proposed in [149, 150], as illustrated in Figure 3.12b. In this case, the LF image is decomposed into its VIs components, which are then KLT coded. It is worth noting that there is no further information on how the resulting KLT coefficients together with the KLT matrix are coded and transmitted in [149, 150]. It is shown that this approach achieves better rate-distortion performance compared to JPEG and the same KLT-based approach applied to MIs. In addition, it is observed that the statistical characteristics between VIs are more easily decorrelated than between MIs, having most of the relevant information compacted into a smaller number of eigenvectors. As stated in [150], this can be explained by two main factors:

1) Due to the small FOV of the microlenses in the array, each captured MI comprises only a small portion of the 3D scene, which may have different characteristics in different areas of the 3D scene. On the contrary, all VIs comprise the complete 3D scene, which are only slightly different on the angles of projection.
2) The MI cross-correlation is differently distributed in a neighborhood depending, for instance, on the distance of the objects relatively to the camera, whereas the orthogonal projection in the VIs removes this depth dependency on the correlation of the objects, and the size of the objects in the VIs is invariant to depth.

3.4.3 DWT-Based Coding

In alternative to block-based transforms, such as the DCT and KLT, some authors proposed to use an approach based on DWT coding, closer to the coding techniques used in JPEG 2000 codecs [112].

In [151], a 3D DWT-based coding scheme is proposed, following the classical still image coding architecture illustrated in Figure 3.2. For this, the LF image is firstly decomposed into a stack of VIs and a separate 1D DWT is recursively applied in the third dimension of this stack until the lowest frequency band contains only two samples (in the third dimension). Then, a two-level 2D DWT decomposition is applied to these two sets of lowest frequency bands. Similarly to JPEG 2000, the lowest frequency subbands are quantized using a deadzone quantizer, while the remaining high frequency coefficients are quantized using a uniform scalar quantizer. Following this, a new scanning pattern is proposed to be used to scan samples from all subbands together, which are then arithmetic coded.

In [152], a similar approach with a 3D DWT applied to a stack of VIs is proposed. However, in this case, the three (separable) 1D DWT are recursively applied to each dimension of the stack, producing 8 subbands in each decomposition level. Afterwards, the 3D DWT coefficients are quantized using a deadzone scalar quantizer and coded using the method of Set Partitioning In Hierarchical Trees (SPIHT). Similarly to the Embedded Block Coding with Optimal Truncation (EBCOT) [153], used in JPEG 2000, the SPIHT algorithm is used as a form of entropy coding applied to bit-planes of quantized coefficients to allow progressive transmission of the LF data. The proposed approach is compared to a 2D version of the coding scheme, in which a 2D DWT is applied to the entire LF image followed by SPIHT. The 3D DWT scheme presents significant improvements compared to the 2D DWT. Moreover, several DWT bank filters are analyzed for the 3D DWT coding, and the Biorthogonal 2.2 filters show the best results, but are very similar to the Daubechies filters.

Regarding standard DWT-based coding solutions, a study on LF image coding is presented in [154], in which the performance of two DWT-based coding solutions (JPEG 2000 and SPIHT) and one DCT-based standard solution (JPEG) are compared. The performance is analyzed in terms of the objective quality of views rendered from the coded and reconstructed LF image, by using average PSNR and average SSIM index. It is shown that the SPIHT scheme presents better RD performance than JPEG for low bitrates, but JPEG 2000 outperforms them both for either PSNR or SSIM metrics. In [155], a similar study is performed for comparing two standard solutions, JPEG 2000 and JPEG XR, for LF image coding. This study focuses on comparing the performance, in terms of objective quality of rendered views, of the different transform coding solutions used in each standard, namely: the JPEG 2000 DWT and the block-based transform used in JPEG XR [156]. In the presented results, JPEG 2000 achieves slightly better RD performance than JPEG XR both in terms of PSNR and SSIM metrics.

In addition, an empirical performance analysis is presented in [92] for computer generated LF images. For this, JPEG 2000 is compared to its extension for volumetric data compression (JPEG 2000 Part 10 [157], known as JP3D) for LF image compression. The JP3D solution supports 3D DWT decompositions and extends tiles, code-blocks and region-of-interest functionalities accordingly to support volumetric data. In the presented study [92], JPEG 2000 is applied to the entire LF image, while two different scenarios are considered for JP3D, in which the 3D DWT is applied to stacks of MIs, and to stacks of VIs. It is shown that the JP3D solution outperforms the JPEG 2000 for both scenarios.

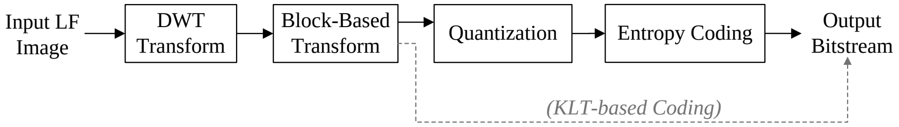


Figure 3.13 Block diagram of combined-transform coding schemes for LF image compression. When using the KLT as the block-based transform, quantization and entropy coding processes are bypassed (as illustrated by the gray dashed line)

3.4.4 Combined-Transform Coding Methods

This category corresponds to LF image coding schemes in which two or more types of transforms are combined to separately exploit the spatial redundancy between samples in a local neighborhood and the inherent LF cross-correlation in neighboring MIs (or VIs). A common characteristic of these LF coding approaches is that an image-based transform (notably, the 2D DWT) is used to exploit the local sample correlation, followed by a block-based transform that is applied to the lowest subbands across different MIs (or VIs), as illustrated in Figure 3.13. The major motivation for this choice is to reduce the blocking artifacts that is likely to appear in the reconstructed image when using a block-based transform coding.

In [158], a combined-transform coding scheme is presented that combines a 2D DWT with a 2D DCT. In this scheme, the LF image is divided into tiles with the MI size to be recursively decomposed with a 2D DWT. Following this, a packet partition scheme is used to rearrange the samples from the same DWT subband into blocks of 8×8 samples to be DCT coded. The 2D DCT coefficients are then scalar quantized and entropy coded similarly to JPEG encoding. The presented approach outperforms JPEG with significant gains, mainly at low bitrates.

In [159], an approach combining a 2D DWT and a 3D DCT is proposed for coding lenticular-based images (see Figure 3.10). For this, the 2D DWT is recursively applied to each VI extracted from an LF image to decompose them in two levels. Then, the lowest subbands from different VIs are stacked into 8×8×8 blocks to be processed by a 3D DCT. Afterwards, all coefficients within all the subbands are quantized using a deadzone scalar quantizer. Then, the 3D DCT coefficients are Huffman coded while all the other coefficients are arithmetic coded. In [160], the previous solution is extended for full parallax LF image, which is then compared to the solution in [139]. It is shown that the combined-transform approach achieves improved RD performance at low bitrates when compared to the scheme proposed in [139], in which only the 3D DCT is used.

In [161], instead of using the DCT as the block-based transform, the 2D DWT is combined with the KLT. In this case, a one level 2D DWT decomposition is individually applied to all MIs, resulting in four subbands per MI. Then, samples in the same subband from all MIs are arranged into four arrays, and the KLT is applied to each of them. Then, the four arrays with reduced dimensionality are transmitted together with the KLT matrix. It is shown that the combined solution performs significantly better, in terms of RD performance, compared to a scheme very similar to the one proposed in [147] (see Figure 3.12b), where only the KLT is used. Moreover, several filter banks are tested for the 2D DWT, and the Daubechies is the one that results in better RD performance for LF image compression.

3.4.5 Extending Transform-Based Methods for LF Video Coding

Only very few schemes [136, 162] propose to extend the transform-based approach to LF video coding, notably, for lenticular-based video content. The idea in these schemes is to combine a 3D DCT, to exploit all the spatial correlations (within the same MIs and also between neighboring MIs), with a differential coding approach, to exploit the existing temporal correlation between adjacent frames. In [136], the Differential Pulse Coding Modulation (DPCM) coding is proposed to be included in a 3D DCT-based coding loop in order to simply encode the difference from a previous frame. In [162], a hybrid coding scheme with motion compensated prediction is proposed, which is based on MPEG-2 [36] but with different block partitioning for motion estimation. To reduce the computational complexity of the motion estimation process, the LF content is firstly decomposed into its VIs and motion estimation is applied only to the central VI. Then, the estimated motion vectors are simply copied for the remaining VIs and this compensated image is subtracted from the original one. The difference is then 3D DCT coded and uniformly quantized. Finally, the quantized coefficients are Huffman coded.

3.5 LF Content Coding Based on Inter-View Prediction

Instead of exploring the inherent LF correlations in the transform domain, as discussed in the previous section, other authors have proposed to explore it in a predictive manner, as identified in Figure 3.1. Notably, some coding schemes propose to extract the MIs/VIs from the LF image in order to represent data as a set of views and to use inter-view prediction for achieving compression.

In this context, this section reviews two types of LF image coding solutions based on inter-view prediction, which are distinguished according to the adopted coding architecture, as shown in Figure 3.14: i) the Pseudo Video Sequence (PVS)-based, in which the MIs/VIs are encoded as a temporal sequence with a 2D video coding standard (Figure 3.14a); and ii) the multiview-based, in which MIs/VIs are encoded as multiview content using a 3D video coding solution (Figure 3.14b).

Although conceptually different (see Figure 3.14), both PVS- and multiview-based coding approaches proposed in the literature have the same basic purpose of proposing an efficient prediction configuration for better exploiting the correlations between the MIs/VIs. For this,

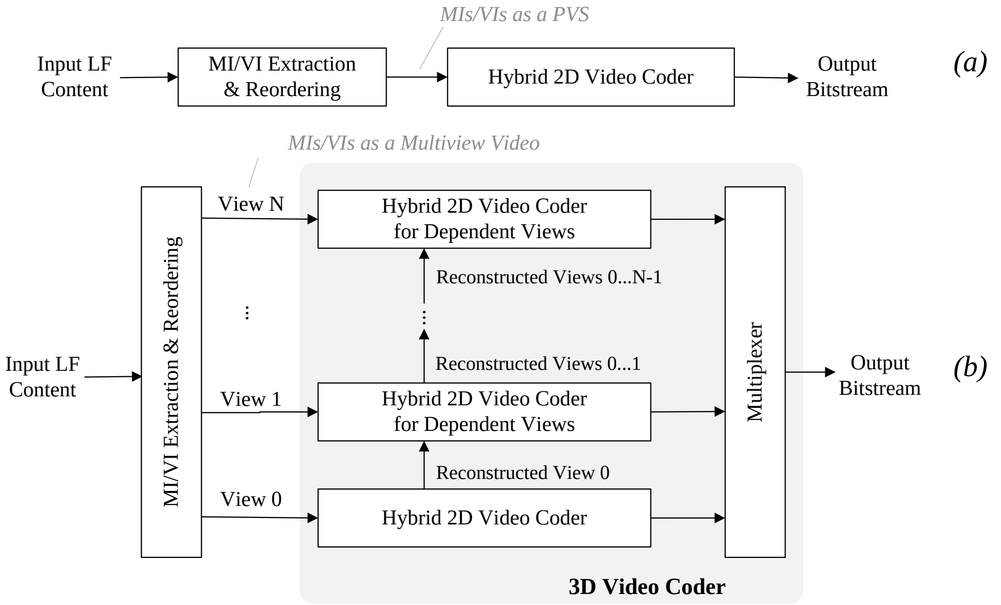

MIs/VIs as a PVS
Input LF Content
MI/VI Extraction & Reordering
Hybrid 2D Video Coder
Output Bitstream
(a)
MIs/VIs as a Multiview Video
Input LF Content
MI/VI Extraction & Reordering
View N
View 1
View 0
Hybrid 2D Video Coder for Dependent Views
Reconstructed Views 0...N-1
Reconstructed Views 0...1
Hybrid 2D Video Coder for Dependent Views
Reconstructed View 0
Hybrid 2D Video Coder
Multiplexer
Output Bitstream
(b)
3D Video Coder

Figure 3.14 LF content coding based on inter-view prediction: (a) PVS-based, and (b) Multiview-based data arrangements

different scanning patterns for ordering them (as summarized in Figure 3.11), as well as different prediction structures (as summarized in Figures 3.15 and 3.16), are proposed.

3.5.1 PVS-Based Image Coding

Back in 1995, the very first LF coding solution in the literature [163] (to the best of the author’s knowledge) proposed to introduce a DPCM coding into a DCT-based image coding loop in order to encode lenticular-based LF images. For this, the LF image was organized as a PVS of MIs and then encoded with the proposed codec, by using the previously encoded MI as the predictor. Since then, the PVS-based approaches proposed in the literature have naturally followed the evolution of hybrid 2D video coding standards.

In [164], a combination of JPEG and MPEG-1 [165] standard codecs is used to encode LF images represented as a PVS of MIs. For this, two prediction configurations are used to exploit the correlations between neighboring MIs, as illustrated in Figure 3.15a. It is shown in [164] that the proposed hybrid coding solution always outperforms the JPEG standard (used for encoding the entire LF image, without MI extraction) with significant gains when the PBI configuration in Figure 3.15a (bottom) is used. In [166], the PVS of MIs is proposed to be encoded with MPEG-2 [36], using the prediction structure shown in Figure 3.15b and three different scanning patterns: i) raster; ii) perpendicular; and iii) spiral (see Figure 3.11). From the presented results, it is shown that the inter-view prediction using raster scan order is much less efficient than spiral and perpendicular due to the reduced number of available vertical inter-view predictions and increased distance between coding frame and reference frame(s). Due to similar reasons, a variation in the inter-view performance is also observed

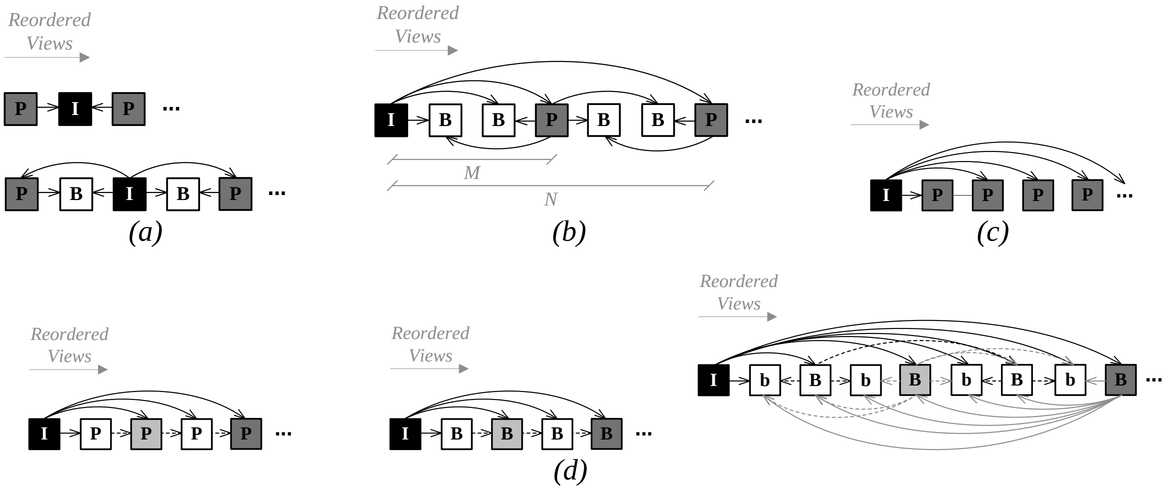


Figure 3.15 Prediction structures proposed in the literature for PVS-based LF coding approaches: (a) PIP (top) and PBI (bottom) [164]*; (b) MPEG-2-based structure with* $M = 3$ *and* $N = 6$ [166]*; (c) Central 1D structure* [167, 168, 173]*; and (d) HEVC-based structures* [175]*: Low Delay P (left), Low Delay B (middle), and Random Access (right), where the B slices in different temporal layers* [123] *are illustrated with different colors and capitalizing format*

for different prediction configurations, i.e., for different M and N parameters in Figure 3.15b. In [105], the authors propose to use a Hilbert scan (see Figure 3.11f) for ordering the MIs in a PVS, which is then encoded using MPEG-2 [36], as in [166]. It is shown that the Hilbert scan results in significant improvement in RD performance compared to the raster, perpendicular, and spiral (see Figure 3.11).

In [167, 168], the authors propose to scan the LF data, represented as a 2D grid of VIs (seen in Chapter 2), in spiral order (Figure 3.11d) and to encode the resulting PVS with a combination of DPCM and the MPEG-4 standard. For this, the prediction structure depicted in Figure 3.15c is used, in which the central VI is considered the only reference frame available for coding all the remaining VIs. Then, the difference between the current VI being coded and its reference frame is encoded as a frame using MPEG-4. The proposed scheme outperforms a JPEG-based solution, in which each VI is independently coded with JPEG. Afterwards, in [169], the authors propose to improve the performance of the solution proposed in [167, 168] by replacing the previous DPCM-based scheme by a motion compensated prediction, and to encode the motion compensated residue using MPEG-4. Moreover, in [170], for further RD performance improvements, the authors propose to equally divide the array of VIs into four parts before scanning them in spiral order to form the PVS, so as to reduce the distance between a current VI being coded and its reference frame.

In [171], the LF data are represented as a PVS, by scanning its VIs in raster order (see Figure 3.11a), which is then encoded with H.264/AVC. The proposed encoder is then compared against JPEG and JPEG 2000, where both standard solutions are used to encode the entire LF image, without VI extraction. From the presented results (for computed generated images using a pinhole lens array model approximation [171]), JPEG 2000 is shown to be more efficient than JPEG for LF image coding. However, the VI-based PVS coding solution outperforms JPEG 2000 with significant gains. In [172], the authors propose to extend their previous work [171] by considering raster, parallel, spiral, and zig-zag (respectively, in Figure 3.11a, b, d, and e) scan for ordering the VIs in the PVS. It is seen that, for computer generated LF images as in [171], changing the scanning patterns does not result in significant differences in the RD performance. Moreover, a comparison between the proposed VI-based PVS and an MI-based PVS (both coded with H.264/AVC) is also performed for LF images generated with different MI and VI sizes. It is suggested that the VI-based PVS shall be preferred to an MI-based PVS when the VI size is larger than the MI size. However, from the presented results, it is also shown that the difference in RD performance between these two PVS approaches (MI- and VI-based) may also depend on the scene characteristics. For instance, VIs extracted from a scene with highly detailed objects distributed in various depth planes are more difficult to encode than MIs, since near and far objects are equally noticeable in the VI [172] (in other words, the objects size is invariant to depth in the VI due to its orthographic property). Differently, in the MIs, near objects are more noticeable than far objects [172] (which are usually blurred in natural LF images). In [92], a performance comparison between a PVS-based approach and a transform-based approach is proposed. For this, the PVS-based approach using H.264/AVC proposed in [172] is compared to a transform-based approach using JP3D standard, in which a 3D DWT is applied to stacks of MIs and VIs. It is shown in the results that the PVS-based approach (using H.264/AVC) outperforms the transform-based approach (using JP3D) with significant gains at lower bitrates, and produces less visible distortions.

Also regarding a comparison against transform-based approaches, a hybrid coding solution combining a motion compensated prediction and a KLT transform coding is proposed in [173], and compared against the KLT-based approach proposed in [149, 150]. For the proposed hybrid solution, a PVS of VIs is extracted from the LF image in spiral scan order, and the central VI is used as the reference frame for encoding all the remaining VIs (as illustrated in Figure 3.15c). Then, the Normalized Cross-Correlation (NCC) [174] is used as the matching criterion for the motion estimation process and, afterwards, the residual information (from motion compensation) is then encoded using the KLT-based coding scheme proposed in [149, 150]. From the experimental results, it is shown that the proposed hybrid solution achieves significant bit savings compared to the transform-based solution in [149, 150].

More recently, in [175], a performance study of HEVC-compatible coding solutions for LF images captured by the Lytro-Illum camera [12] is presented. For this, the used Lytro-Illum LF images are firstly pre-processed with the LF Toolbox version 0.4 [85], resulting in a rectified LF image in which the original hexagonal grid of MIs is transformed to a square grid. From this rectified LF image, a VI-based PVS using two different scan orders – raster and spiral (see Figure 3.11) – is encoded with HEVC using three different prediction structures – Low Delay P, Low Delay B, and Random Access [176] (see Figure 3.15d). The different PVS-based approaches are also compared to the case where the rectified LF image is encoded in its entirety using HEVC Still Picture Profile [2]. From this study, it is shown that not only the different scanning order and prediction structures highly influence the achieved compression efficiency, but also the characteristics of the captured scene. In other words, the relative RD performance of the compared coding approaches is not consistent for all images, and, consequently, it is not possible to reach a general conclusion regarding which would be the best representation format for encoding Lytro-Illum LF images.

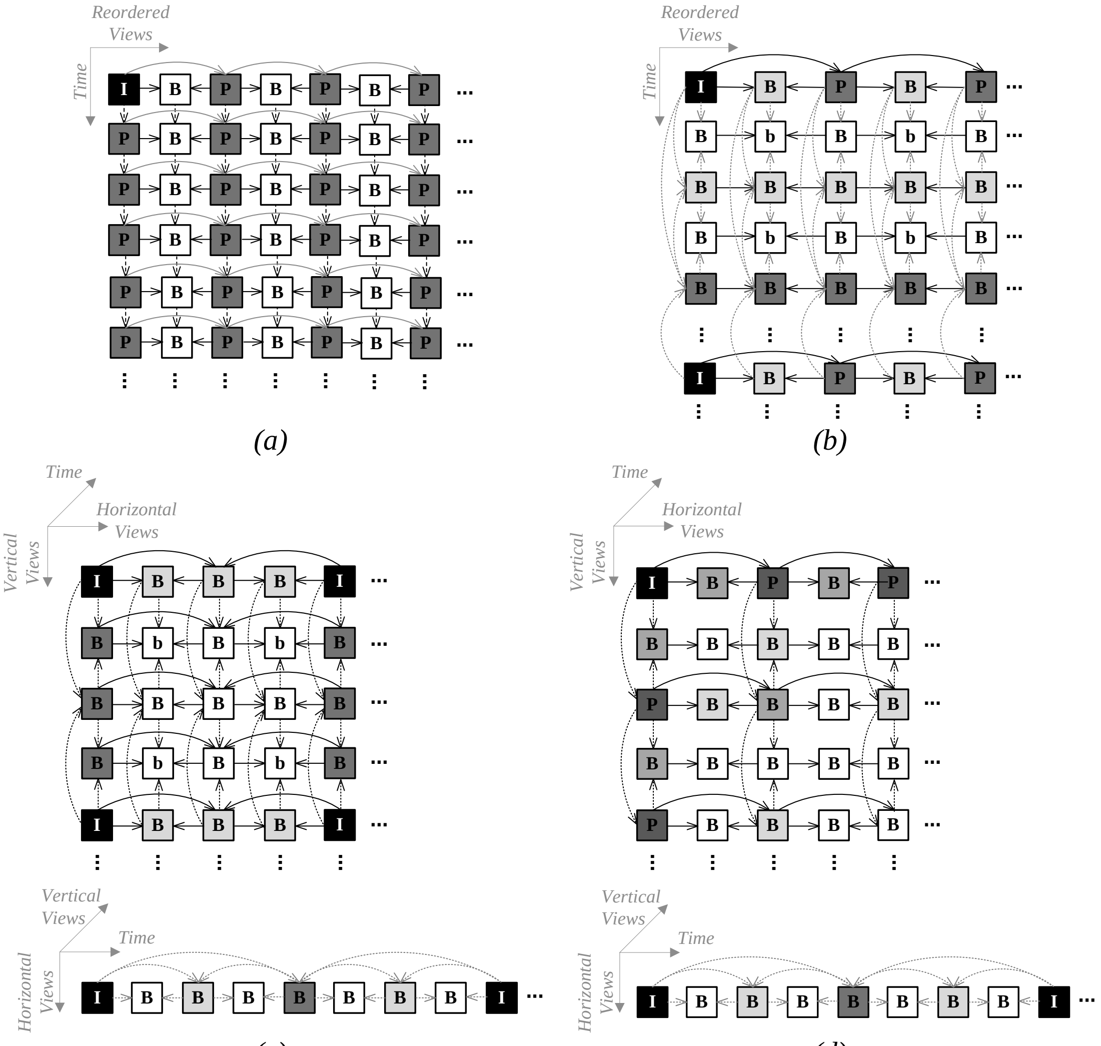
Although originally proposed in the literature for LF images based on camera arrays in [177], LF video can also be coded with a PVS-based approach by using the called transposed picture ordering [177]. In this case, all views from the same time instant are concatenated along the time dimension. However, it is worthwhile to note that the temporal correlation between adjacent time instants no longer exists in the final video sequence.

3.5.2 Multiview-Based Video Coding

The multiview-based video coding approaches can be seen as an alternative to the abovementioned transposed picture ordering [177] in which the representation format and prediction structure of standard 3D video coding solutions – such as MVC [39], MV-HEVC, and 3D-HEVC [40] – are used for coding LF video arranged as multiple MI/VI sequences. Therefore, the basic idea of the coding approaches in this group is to adapt the conventional inter-view prediction structure for full parallax in order to improve the RD performance for LF video coding.

In [178, 179], the authors propose to decompose the LF video into multiple VI video sequences scanned in raster order (see Figure 3.11) and to jointly exploit motion (temporal prediction) and disparity (inter-view prediction) similarly to what is done in MVC [39]. For this, the prediction structure depicted in Figure 3.16a is adopted, and the Evolutionary Strategy (ES) is used to speed up the motion estimation process. In [180], the authors significantly improve their previous work by using motion estimation with half pixel precision. In [181], the authors propose to scan the VI video sequences in raster order and to encode it with MVC by using the conventional prediction structure used in MVC as shown in Figure 3.16b. It is seen that the proposed solution outperforms a H.264/AVC-based coding solution, in which the LF video sequence is encoded in its entirety using H.264/AVC. It is worthwhile mentioning that these coding schemes [178–181] consider only lenticular-based content (i.e., only horizontal parallax) with a small number of VIs (up to eight).

In [182], the 2D grid of VI sequences is firstly scanned in spiral order to obtain a horizontal arrangement of VI sequences, which are then encoded with MVC using the same prediction structure used in [181], as illustrated in Figure 3.16b. From the presented results (for a computer generated LF video with 3×3 VI sequences), the multiview-based arrangement



(a) *(b)*

(c) *(d)*

Figure 3.16 Prediction Structures proposed in the literature for mutiview-based LF coding approaches: (a) IBP structure [178, 179]*; (b) Based on the typical prediction structure used in MVC ; (c)2D hierarchical inter-view prediction structure proposed in* [183]*; and (d) 2D parallel inter-view prediction structure proposed in* [184]

outperforms a PVS-based approach (in which the VI sequences are reordered using transposed picture ordering [177] and encoded with H.264/AVC) as well as a H.264/AVC-based coding solution (in which the LF video sequence is encoded in its entirety using H.264/AVC) with significant RD gains.

In [183], a hierarchical 2D inter-view prediction structure, shown in Figure 3.16c is proposed for LF video coding using MVC. The idea is to optimize the inter-view prediction structure to the 2D grid of VI sequences, and then to further minimize the distance between the current VI and their inter-view reference frame(s). The proposed hierarchical prediction structure is compared against the Hilbert scan of VIs proposed in [105], and presents expressive RD performance improvements. Moreover, a parallel implementation of the proposed prediction structure is also designed, which significantly reduces the overall encoding time. Similarly, in [184], a 2D parallel inter-view prediction structure, shown in Figure 3.16d, is proposed for coding LF video using MVC. The proposed prediction structure is compared against the conventional MVC prediction structure (corresponding to the solution proposed in [181]) and against a spiral scan of VIs proposed in [185], in which the central VI is considered the only inter-view reference for all the remaining VIs (similar to the prediction structure shown in Figure 3.15c). From the presented results (for video sequences captured using the Lytro first generation camera [12]), it is seen that the proposed solution outperformed the other two tested solutions (except at low bitrate values, where it is outperformed by the MVC prediction structure).

An advantage of using a standard 3D video coding solution is that scalability and backward compatibility are straightforwardly supported. However, the texture resampling from micro-images to the viewpoint images usually results in very low resolution images with significant aliasing artifacts [86]. Motivated by this fact, an alternative to the multiview-based data arrangement using VIs has been proposed in the context of this Thesis, which is going to be presented in Chapter 5.

3.6 LF Content Coding Based on Non-Local Spatial Prediction

As an alternative to inter-view prediction (see Figure 3.1), some LF coding solutions proposed to exploit the non-local spatial correlation that exists between MIs in the LF image to achieve compression in a predictive manner. For this, the LF image is encoded and transmitted as a 2D grid of MIs (i.e., using the raw MI-based 2D format seen in Chapter 2) that is then encoded with a hybrid 2D video codec and by using a special LF non-local spatial prediction, as depicted in Figure 3.17.

This LF coding approach can be then grouped into two types of LF non-local spatial prediction schemes that will be explained in the following: i) spatial displacement compensated prediction [186, 187]; and ii) neighbor-embedding prediction [188].

3.6.1 Spatial Displacement Compensated Prediction

The idea of exploiting non-local spatial redundancy has been firstly proposed for 2D image and video compression in order to further enhance the performance of H.264/AVC intra prediction. Notably, the intra macroblock compensation technique [186, 187] proposes to extend the usage of motion compensated prediction for intra-frames in order to reduce the number of bits needed in conventional intra coding while still supporting random access [187].

LF image compression using non-local spatial prediction has been firstly proposed in the context of this Thesis in [60, 61] with the purpose of exploiting the existing MI cross-correlation of this type of content. For this, the SS compensated prediction method is

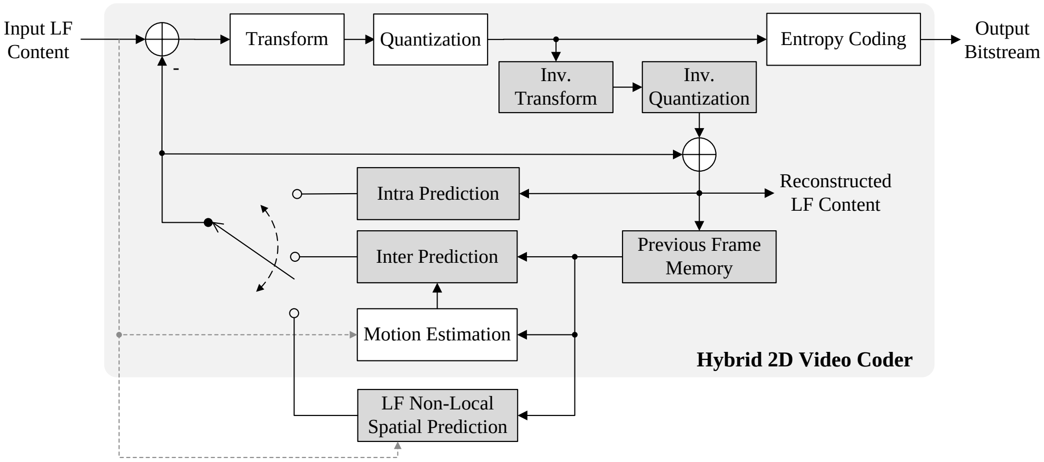


Figure 3.17 LF image and video coding based on LF non-local spatial prediction

proposed to improve the performance of H.264/AVC standard for LF image coding. Similarly to motion compensation, the SS estimation process uses a block-based matching over a causal search window (i.e., a search window containing only previously coded pixels, as seen in Figure 3.18a), to find the 'best' predictor (in an RDO sense) for the current block. As a result, the chosen block becomes the best candidate predictor and the relative position between the two blocks is signaled by a displacement vector (see Figure 3.18), referred to as SS vector. The SS compensated prediction method extended the idea of [186, 187] by including additionally an efficient vector prediction scheme, as well as by specifying different ways to partition the macroblock for employing the SS estimation and compensation. In [62], the SS prediction concept is further extended for inter-coded frames based on H.264/AVC. It is shown that the LF video coding RD performance can be improved by exploiting the existing correlations in all domains (i.e., within the same MI, across neighboring MIs and across adjacent temporal frames). Also in the context of this Thesis [53–55] (see Chapter 4), the SS prediction is proposed to be added to the HEVC coding architecture for image and video coding, so as to take advantage of the flexible partition patterns used in this type of video codecs. It is shown in [54] that, independently of the codec technology (namely, H.264/AVC or HEVC), the SS prediction technique further improves the RD performance relatively to the original version of the codec. Moreover, the obtained RD gains are less dependent on the underlying codec and more related to the characteristics of the LF content. The complete proposal of an LF coding solution based on HEVC and with SS compensated prediction is going to be presented in Chapter 4.

Subsequently, in [189], the authors propose to extend the SS prediction concept by using HEVC inter B frame bi-prediction to further improve the RD performance for LF image coding. However, in this case, to guarantee that the two prediction signals come from two different MIs, the search area is separated into two non-overlapping parts [189], which are assumed to be two different reference pictures, one in each of the two reference picture Lists

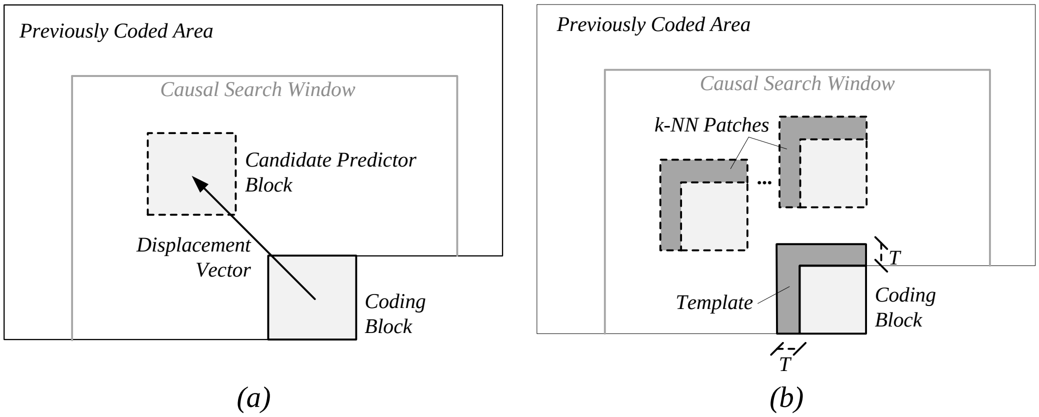


Figure 3.18 Non-local spatial prediction approaches: (a) Spatial displacement compensated prediction; and (b) Neighbor-embedding prediction

0 and 1 [189]. Then, the best uni-predicted candidate predictor from each reference picture list is found. Therefore, as in HEVC inter B frame coding, three candidate predictor are derived and the best among them is chosen in an RDO sense: i) the best uni-predicted candidate from List 0; ii) the best uni-predicted candidate from List 1; and iii) the linear combination of the previews candidates (i and ii) for bi-prediction.

Although not targeting LF image coding, a very similar scheme, known as Intra Block Copy (IntraBC) [190], has been recently proposed in the context of HEVC Format Range Extension (RExt) [191] and HEVC Screen Content Coding (SCC) [190] standardization developments. Firstly proposed in [192], it aims at improving the HEVC coding efficiency for screen video compression, motivated by the fact that this kind of content often contains a substantial amount of still or moving rendered graphics and texts with repetitive patterns. Similarly to the schemes in [53, 189], the IntraBC also uses a block-based matching algorithm to estimate a displacement vector that indicates the relative position of the predictor block to the current block being coded. However, the estimated vector uses only integer pixel accuracy. Moreover, in the early developments of HEVC RExt standard [193], the IntraBC prediction was allowed for a smaller set of PU partition patterns and limited CB sizes, and only 1D displacement vectors are estimated across one or more CBs to the left. More recently, an improved IntraBC version, proposed in [194, 195], is adopted as part of the SCC reference software. In this case, the search window is expanded to the entire CB row or column (for 16×16 CBs), or to the entire previously coded area of the picture by using a hash-based search (for 8×8 CBs). It is worth noting that RExt was included in the second version of the HEVC standard (finalized in October 2014), and SCC was still under development at the time of writing this Thesis.

3.6.2 Neighbor-Embedding Prediction

Neighbor-embedding prediction methods have been increasingly considered for still image compression [188], being especially powerful for predicting highly complex textured areas of the image. The idea of this type of prediction is basically to search for an optimized combination of *k*-NN texture patches that best approximate known sample values in a previously coded neighborhood of the coding block (known as template), as depicted in Figure 3.18b. Then, the same combination of coefficients is used to approximate the unknown samples in the coding block (see Figure 3.18b). Examples of neighbor-embedding prediction methods proposed in the literature for image coding are the Non-negative Matrix Factorization (NMF) [196] and the LLE [197] dimensionality reduction techniques. Moreover, the Template Matching (TM) [198] algorithm can be seen as a particular case of neighbor-embedding prediction, in which a unique nearest neighbor, 1-NN, texture patch is found with the linear weighting coefficient equal to 1 [188].

With regard to LF image compression, the work in [199] proposes to replace one of the conventional intra directional prediction modes of HEVC by a prediction scheme based on TM for better adapting to the repetitive MI texture patterns. Similarly to TM, the proposed prediction method uses an implicit approach to avoid transmitting any information about the used predictor. In this scheme, three neighboring CBs are used as the template and two separate search windows are adopted for finding two best predictors to this template. The first search window comprises all candidates with the same shape and in the same row as the template (horizontal search window), and the second comprises all candidates in the same column of the template (vertical search window). From the selected predictors, two 1D vectors are derived, and their combination determines the block predictor for the current CB. It is shown that the proposed scheme outperforms the TM algorithm (when including the TM in the HEVC coding framework) for LF image coding, being able also to considerably reduce the computational complexity for the same search window.

In [70], a neighbor-embedding prediction method based on LLE is proposed for LF image coding based on HEVC. In this case, the predictor to the current CB is given as a linear combination of its *k*-NN patches inside a causal search window in the LF image (see Figure 3.18b). To avoid transmitting information about the selected *k*-NN patches, the LLE method searches for them, in terms of Euclidian distance, at both encoder and decoder side. Afterwards, the set of weighting coefficients for combining the *k*-NN patches are determined by solving a least-squares optimization problem with a constraint on the sum of the coefficients that has to be 1. After finding the optimally estimated coefficients, the predictor block is determined by using the same linear coefficients estimated for the template to combine the square blocks associated to each *k*-NN patch (see Figure 3.18b). For improved performance, the encoder tests different *k* values, from 1 up to 8, and the one that produces the best block prediction result (in RDO manner) is explicitly transmitted to the decoder. To avoid further signaling in the HEVC bitstream, up to 8 HEVC intra directional modes are replaced by the LLE-based mode using a different number of *k*-NN patches. From the presented results, it is shown that the proposed LLE-based coding solution always outperformed HEVC Still Image Profile (with bit savings of up to 38 %) and the SS-based solution proposed in [53, 54] (up to 15 % of bit savings when the number of k-NN templates is adaptively chosen by varying from 1 to 8).

More recently, in [200], the authors proposed an HEVC-based neighbor-embedding predictive solution using Gaussian Process Regression (GPR) for LF image coding. For this, k-NN patches are firstly chosen in terms of Euclidian distance in a causal search window similar to [70]. However, in order to reduce the computational complexity, the causal search window is divided into two different search windows (horizontal and vertical search windows, similarly to what is done in [199]), and the template thickness, T (Figure 3.18b), is substantially reduced. Then, a filtering method based on the NCC [174] is used to judge the reliability of the obtained k-NN patches. Afterwards, the prediction from the k-NN patches is modeled as a non-linear (Gaussian) process, and GPR is then used for estimating the predictor block. The GPR-based neighbor-embedding prediction is then included into HEVC by replacing one of the HEVC intra directional modes and no further signaling is needed. The proposed GPR-based is compared against HEVC SCC [190], as well as against the TM-based solution proposed in [199] and the LLE-based solution proposed in [70] (but fixing 6-NN patches, instead of adaptively varying k from 1 to 8 as in [70]). It is shown that the proposed GPR-based solution outperforms the TM-based solution [199] and HEVC SCC (with up to 21 % of bit savings) with significant coding gains, and always outperforms the LLE-based solution (with up to 5 % of bit savings), showing that improved prediction results could be obtained by using GPR instead of LLE for texture and edge regions.

3.7 Disparity-Assisted LF Image Coding

Mainly motivated by the design of recent 3D coding solutions (such as 3D-HEVC), disparity-assisted LF coding approaches have been recently proposed in the literature for LF image coding.

Differently from the previously mentioned predictive LF coding approaches, the coding solutions in this category aim at greatly reducing the amount of LF texture data that is encoded and transmitted in order to achieve compression. For this, a particular subsampled texture representation of the LF data is proposed to be used, which is then encoded and transmitted together with the corresponding disparity information. Hence, at the decoder side, the subsampled texture plus disparity is used to estimate the information discarded at the encoder side. Since this reconstructed LF image is, in some cases, used as a reference frame for prediction, the disparity-assisted approach is also connected to the node of predictive LF coding solutions as depicted in Figure 3.1.

From the proposed solutions, two types of subsampled LF representations can be recognized, namely: i) sparse set of MIs; and ii) sparse set of (rendered) views. LF image coding solutions in each of these two types are thus reviewed in this section.

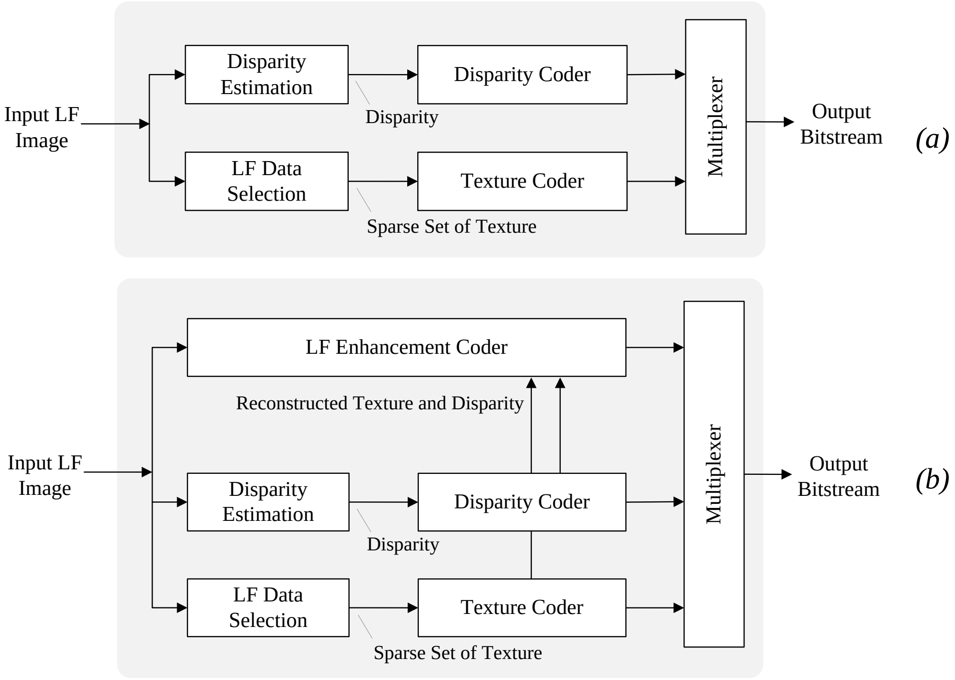


Figure 3.19 Disparity-assisted LF image coding: (a) Basic coding architecture proposed in [204, 205], *and (b) Alternative coding architecture proposed in* [206, 207]

3.7.1 Sparse Set of MIs plus Disparity

In [201–203], the authors have firstly proposed to represent the LF data by a sparse set of MIs that are uniformly subsampled from the LF image in order to remove the redundancy between neighboring MIs and to achieve compression. It is shown that it is possible to reconstruct the discarded MIs at the decoder side by simply using the coded and reconstructed MIs together with information about the optical geometry used when acquiring the LF content. Moreover, the proposed scheme is able to improve the RD compression performance when incorporated into the JPEG standard [201–203].

Alternatively, disparity-assisted coding schemes are proposed in [204, 205], in which the disparity between adjacent MIs is used to better reconstruct discarded MIs in the sparse set of MIs. Therefore, in these cases [204, 205], the LF image is encoded using the coding architecture illustrated in Figure 3.19a. Notably, in [204], JPEG is used as the texture coder and lossless arithmetic coding is used for the disparity data. In [205], the sparse set of MIs is represented as multiview content and each MI is then encoded using a method similar to MVC simulcast coding [39]. Moreover, the disparity is losslessly encoded using a run-length coding scheme followed by Huffman coding.

At the decoder side, the discarded MIs can then be reconstructed by simply applying a disparity shift (in [204]) or by using a DIBR algorithm modified to support the multiple MIs as input views (in [205]), followed by an inpainting algorithm to fill in the missing areas. However, although high compression can be achieved by this approach, the disparity/depth information is usually estimated from the captured LF image, which, in real-world images, introduces inaccuracies. For instance, in [204], the disparity is assumed to be the same for all pixel positions inside an MI. On one hand, this can be a valid approximation since each MI has a small FOV. On the other hand, this assumption is likely to be inaccurate at object boundaries since a single MI can still capture (small) portions of objects in different depth planes. Hence, the quality of the reconstructed MIs – and, consequently, the quality of rendered views – is severely affected by these disparity inaccuracies at the encoder side. For this reason, instead of uniformly selecting the MIs in the LF image, the selection is carried out adaptively, so as to obtain better view reconstruction [204, 205]. In [204], an iterative selection of MIs is performed based on a cumulative disparity metric. In [205], a visibility test is used to select extra MIs to be encoded and transmitted by identifying possible hole-causing regions.

However, a common characteristic of these approaches is that the quality of rendered views is negatively affected by inaccuracies in the synthesis of the missing MIs. For instance, due to occlusion problems and quantization errors when lossy encoding this disparity/depth, some synthesized MIs may present too many missing areas to be filled, thus introducing even further inaccuracies. The reconstruction artifacts are even more challenging for synthesizing MIs due to their small FOV and resolution (compared to view synthesis in conventional depth-assisted 3D coding solutions). For this reason, in [206], the entire LF image is also encoded and transmitted in an LF enhancement layer, as depicted in Figure 3.19b, so as to provide better rendered views (i.e., rendered from the content in the LF enhancement layer). In this case, an HEVC-based coding scheme is used for (lossy) encoding the sparse grid of MIs (in the base layer), as well as the disparity information represented as 2D images. Then, the reconstructed texture and disparity are used for reconstructing the LF image, which is later used as a reference frame for coding in the LF enhancement layer. From the experimental results, it is seen that the proposed disparity-assisted solution presents significant bit savings (up to 65 % when subsampling the grid of MIs by a factor of 2) compared to encoding the entire LF image with HEVC Still Picture Profile. However, a substantial difference in objective quality between the reconstructed LF content in the lower layers and the LF enhancement layer content is also observed.

3.7.2 Sparse Set of Views plus Disparity

An alternative disparity-assisted coding approach is proposed in [207] also using the coding architecture shown in Figure 3.19b. However, in this case, a sparse set of views is rendered from the LF image and then encoded together with the disparity information. It is worthwhile to notice that, in this particular case [207], the goal is not to use the disparity information to reduce the amount of texture information that is encoded and transmitted, but instead to take advantage of the view rendering process to construct an efficient reference picture for encoding the entire LF image in the LF enhancement layer [207]. For estimating the disparity, the block-based matching algorithm proposed in [10] is adopted, in which a single

4-bit value of disparity is computed for each MI. Then, this disparity information is used to render a single view from the LF image by using the disparity-assisted weighted blending algorithm proposed in [10]. For encoding the rendered view, the 3D-HEVC standard is used (the texture coder in Figure 3.19b), where the coding configuration (notably, the QP value) is selected so as to optimize the RD coding performance in the LF enhancement layer encoding process. On the other hand, disparity information is directly transmitted to the decoder side, bypassing the disparity coder block in Figure 3.19b. Afterwards, at the LF enhancement layer coder (see Figure 3.19b), the reconstructed view is low-pass filtered using an average filter [10]. This filtered view and disparity information are then used to build a reference picture, which is simply subtracted from the original LF image and encoded with HEVC intra coding. From the presented results, it is seen that further bit savings could be achieved (up to 31.1 % of bit savings compared to HEVC Still Picture Profile [2]) by using the proposed approach when an optimized set of QP values are selected. More recently, in [208], the authors proposed to extend their previous coding solution proposed in [10] so as to consider more than one extracted view (notably, a set of 3, 5, and 9 views). These views are then encoded with 3D-HEVC using an IPP inter-view prediction structure and a set of optimized QP values. The proposed solution achieved up to 29.1 % (when 3 views are extracted), 27.9 % (when 5 views are extracted), and 27.2 % (when 9 views are extracted) of bit savings compared to HEVC Still Picture Profile [2].

3.8 Conclusions

This chapter reviewed the most relevant LF image and video coding approaches proposed in the literature. In addition, the chapter also reviewed the relevant state-of-the-art 2D and 3D coding standards and introduced some basic concepts that were relevant for characterizing the variety of LF coding approaches existing in the literature. From this, the LF coding approaches were categorized into four groups depending on how the inherent LF content correlations were exploited. Notably, these were based on: i) transform coding; ii) inter-view prediction; iii) non-local spatial prediction; and iv) disparity-assisted coding.

From the presented overview, it is worthwhile to highlight the following conclusions:

- **Regarding State-of-The-Art Coding Standards** – HEVC was able to provide significantly improved compression performance for both image and video coding relative to its predecessor state-of-the-art image and video coding standards. Moreover, many authors studied the performance of LF content coding using different standard solutions, and similar conclusions were also demonstrated for LF image coding, where HEVC presented significantly better performance compared to JPEG, JPEG 2000, and H.264/AVC.

- **Regarding Transform-Based LF Coding Approaches** – Several LF content coding schemes proposed to exploit the inherent correlations of the LF content in the transform domain. From the different transform coding techniques that were proposed, it was suggested that LF content coding using a DWT coding presents the best RD coding performance, either when applied alone, or combined with a block-based transform – such as DCT or KLT.
- **Regarding LF Coding Based on Inter-View Prediction** – Many authors proposed to extract the MIs and VIs from the LF data to be represented as a PVS or as multiview content, which was then encoded, respectively, with a 2D or 3D content coding solution. When compared to transform-based approaches, these PVS- and multiview-based approaches showed significant RD performance improvements for LF image coding. Moreover, it was seen that these coding approaches leave open the possibility of a huge variety of data arrangements and prediction structures for better exploiting the correlations between MIs/VIs. However, it was suggested that the achieved RD compression efficiency can be highly influenced not only by these factors, but also by the characteristics of the captured scene.
- **Regarding LF Coding Based on Non-Local Spatial Prediction** – Other authors proposed to encode and transmit the LF image as a whole, by using a standard 2D coding solution and a special non-local spatial prediction. The advantage of such systems – when compared to transform-based, PVS-based, and multiview-based approaches – was that the inherent correlation of the LF data could be exploited without any knowledge of the optical geometry used for capturing the LF images. Moreover, it was suggested that efficient non-local spatial prediction schemes were able to significantly improve the performance of standard coding solutions – such as H.264/AVC and HEVC – for LF image and video coding. In addition, further RD improvements for LF coding were also possible regarding other conventional non-local spatial prediction schemes – such as TM and IntraBC.
- **Regarding Disparity-Assisted LF Coding Approaches** – Some authors proposed to incorporate disparity information into the LF coding bitstream in order to greatly reduce the texture information that was encoded and transmitted (together with corresponding disparity). Some promising approaches proposed in the literature showed that it was possible to significantly improve the RD performance compared to H.264/AVC and HEVC standards. However, a limitation of these approaches was that the quality of views rendered from the sparse set of texture plus disparity was negatively affected by inaccuracies in the estimated disparity at the encoder side.

Chapter 4

HEVC-Based Light Field Coding with Self-Similarity Compensated Prediction

Providing an efficient coding scheme to deal with the large amount of data involved in LF systems is a requirement of utmost importance in order to deliver to the end-users LF content with convenient viewing resolution as well as with more powerful capabilities (e.g., in terms of content manipulation). Hence, this chapter will be focused on proposing an efficient Light Field Coding (LFC) solution.

As discussed in the previous chapter, the LF image can be simply represented in a (raw) 2D image, which can be then encoded in its entirety (without decomposing it into sets of MIs or VIs), for instance, using a conventional 2D coding standard. Moreover, it has been seen in Chapter 2 that, as a result of the used optical LF acquisition setup, the planar light intensity distribution in the LF content presents a repetitive structure of MIs that may be exploited for achieving compression with respect to a standard 2D coding solution.

Therefore, this chapter proposes an efficient LF coding solution based on HEVC and using a non-local spatial prediction scheme, named SS compensated prediction, to exploit this inherent redundancy that exists between micro-images in the LF content. For this, the proposed SS compensated prediction makes used of the generic concept of superimposed compensated prediction [51], in which one or more candidate blocks can be estimated (according to appropriate criteria) from the same reference picture, and can be then used to predict the current block being coded. For this, the previously coded and reconstructed area of the current picture itself is seen as a reference frame, referred to as SS reference. Thus, similar to motion estimation, a block-based matching algorithm is used to estimate (inside a search window in the SS reference) the ‘best’ predictor block, in an RDO sense [52], for the current block. This predictor block can be generated from a single candidate block, referred to as the Uni-predicted Self-Similarity (Uni-SS), or from a combination of two different candidate blocks, referred to as the Bi-predicted Self-Similarity (Bi-SS). As a result of this SS compensated prediction, one or two SS vectors are derived, which indicate the relative horizontal and vertical positions of the chosen candidate block(s) with respect to the position of the current block. To take advantage of the distinctive characteristics of these SS vectors, a novel SS vector prediction scheme is proposed in order to achieve further bit savings to the

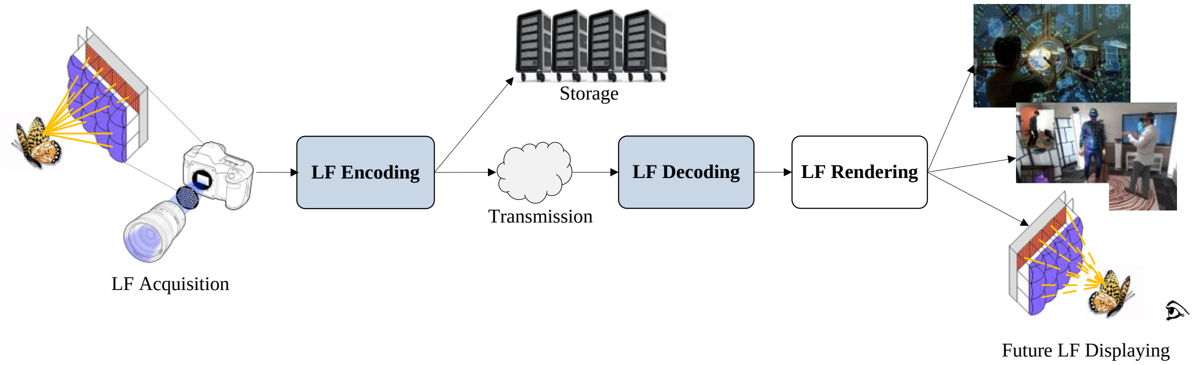


Figure 4.1 Target application workflow for the proposed LF coding solution

proposed LFC solution. Moreover, in the case of the Bi-SS prediction, the two candidate blocks are here proposed to be jointly estimated by using a locally optimal rate-constrained algorithm [58] in order to further improve the RD coding performance of the proposed LFC solution.

The LFC solution proposed in this chapter is not tuned for any particular optical acquisition setup since it does not require any explicit knowledge about it (e.g., microlens size, focal length, and distance of the microlenses to the image sensor). Notice that, although these parameters may be provided by camera makers, many of them are highly dependent on the manufacturing process, being different even from camera to camera of the same model (e.g., each microlens may vary slightly in shape, size, and relative layout position). This means that the LF content from each specific LF camera is also different, and compression tools that use this kind of information need to be robust to these variations. For this reason, using compression tools that are less dependent on a very precise calibration pre-process may be advantageous for supporting LF visualization without increasing the processing complexity.

In this sense, the proposed LFC solution may be mainly advantageous for applications in which the LF content is consumed by the end-user in a format similar to the captured format, as for example, in the case in which the captured LF content is visualized in an LF display that also makes use of an MLA in its optical system, or is consumed by using a proprietary LF rendering algorithm that makes used of the same (raw) 2D format. In this scenario (see Figure 4.1), the captured LF image can be encoded with the LFC solution proposed in this Thesis and may be then transmitted to a receiver over a heterogeneous network. Alternatively, this LFC solution may be also advantageous for improving the storage compression efficiency (see Figure 4.1).

The remainder of this chapter is organized as follows. Section 4.1 presents the proposed LFC codec architecture based on HEVC and using the SS compensated prediction, as well as some preliminary RD performance results with respect to HEVC. Section 4.2 briefly discusses the possibility of extending the proposed LFC solution for LF video coding. Section 4.3 proposes an efficient SS vector prediction to further improve the RD performance

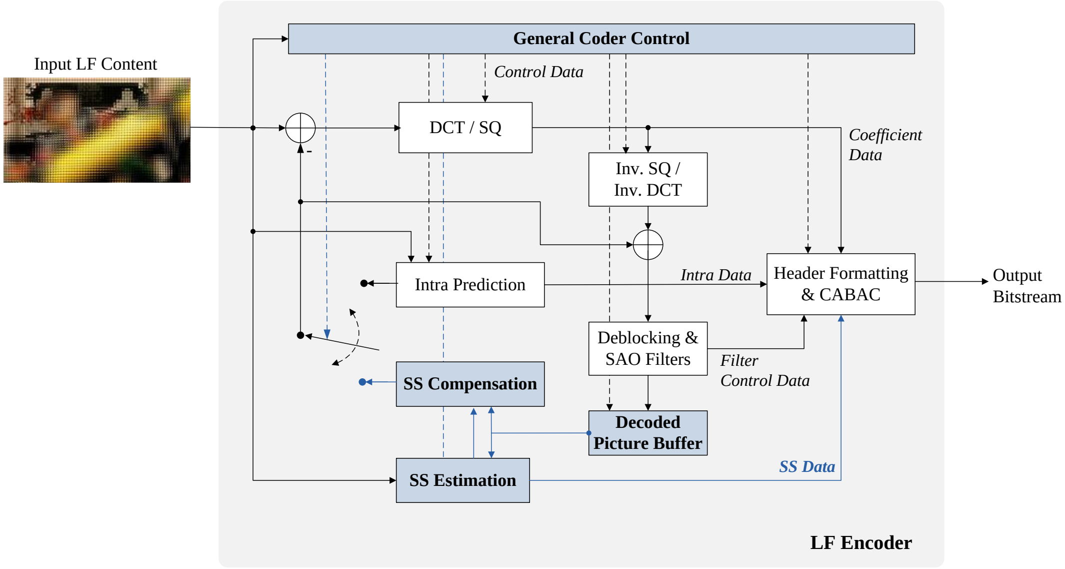


Figure 4.2 Proposed LFC coding architecture with SS prediction (novel and modified blocks are highlighted in blue)

in the proposed LFC solution. Section 4.4 presents the proposed Bi-SS prediction, including a theoretical analysis of the coding efficiency improvements achieved, and an experimental analysis of the SS vector prediction performance for bi-prediction. Section 4.5 presents test conditions and experimental results for evaluating the complete LFC Bi-SS coding solution; and, finally, Section 4.6 concludes the chapter.

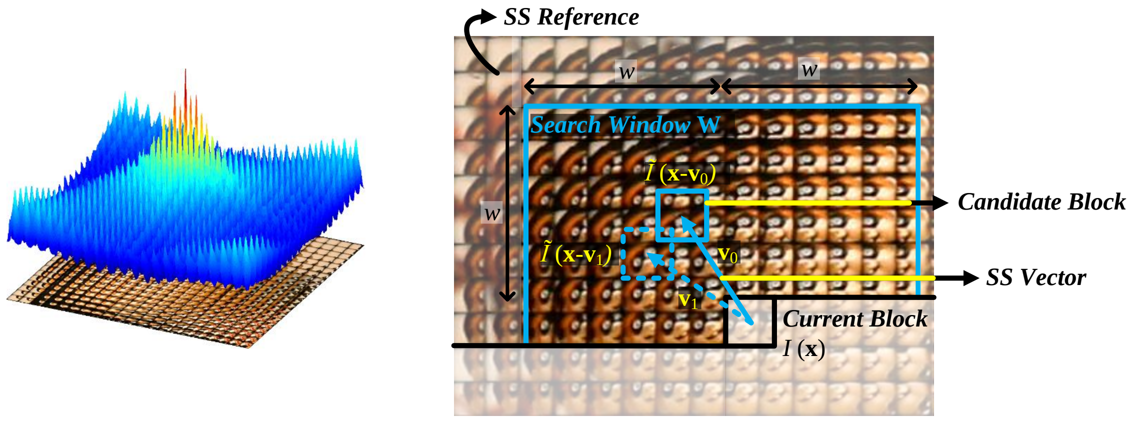
4.1 Proposal of an HEVC-Based LFC Solution Architecture

Figure 4.2 presents the architecture of the proposed LFC solution, which is based on HEVC and comprises both additional and modified modules to efficiently handle LF images. Specifically, the proposed codec introduces an additional type of prediction – the SS – and the encoder will choose the best, among SS and conventional HEVC intra prediction in an RDO manner (as shown in Section 3.1.2).

The novel and modified blocks of the proposed LFC solution (highlighted in Figure 4.2) are explained in this section. In addition, to illustrate the advantage of adding the proposed SS compensated prediction to HEVC for LF image compression, some preliminary experimental results are also presented for Uni-SS prediction.

4.1.1 SS Estimation and Compensation

Enhancing the HEVC coding architecture with SS compensated prediction requires adaptations at the following stages of the coding process (which are managed by the SS estimation and SS compensation blocks in Figure 4.2): i) SS estimation; ii) Prediction modes and block partitioning; iii) SS compensation; and iv) SS vector prediction.



(a) *(b)*

Figure 4.3 The SS compensated prediction concept: (a) Inherent MI cross-correlation in an LF image neighborhood; and (b) The SS estimation process (example of a second candidate block and SS vector for bi-prediction are shown in dashed blue line)

4.1.1.1 SS Estimation

Basically, the SS compensated prediction exploits the cross-correlation existing in an MI neighborhood (see Figure 4.3a) by estimating 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 (referred to as SS reference, as seen in Figure 4.3). Hence, the relative position between the current and the 'best' candidate block is signaled by an SS vector, $\mathbf{v}_0$, (see Figure 4.3b). Similarly to the conventional HEVC inter P frame prediction, the best SS vector, $\mathbf{v}_0^{best}$, for the SS prediction is found by minimizing the Lagrangian cost function in (4.1) [52],

$$\mathbf{v}_0^{best} = \arg\min_{\mathbf{v}_0} \left\| I(\mathbf{x}) - \tilde{I}(\mathbf{x}-\mathbf{v}_0) \right\|_1 + \lambda R(\mathbf{v}_0) \tag{4.1}$$

where $I(\mathbf{x})$ is a matrix variable representing the current block at position $\mathbf{x} = (x,y)$ in the LF image; $\tilde{I}(\mathbf{x}\text{-}\mathbf{v}_0)$ represents a candidate block in the SS reference, $\tilde{I}$, with $\mathbf{x}\text{-}\mathbf{v}_0 \in \mathbf{W}$ (see Figure 4.3b); and $R(\mathbf{v}_0)$ corresponds to the estimated number of bits for encoding the SS vector $\mathbf{v}_0$. In addition, to keep the complexity low, the l_1-norm (or SAD), $\|\ \|_1$, is used and a limited causal search window $\mathbf{W}$ is adopted. However, it is worth noting that the search area shall be larger than the MI resolution so as to be able to exploit the inherent MI cross-correlations. Finally, the SS predictor block is derived as $\tilde{I}(\mathbf{x}\text{-}\mathbf{v}_0^{best})$.

Notice that the SS estimation process in (4.1) only considers a single compensated signal for prediction of the current block, and for this reason will be hereinafter referred to as Uni-SS prediction. Moreover, further improvements to this SS compensated prediction concept will be proposed in Section 4.4 for allowing jointly estimated Bi-SS prediction.

4.1.1.2 Prediction Modes and Block Partitioning

The SS compensated prediction is evaluated for all CB sizes (i.e., from 64×64 down to 8×8) in the conventional RDO process to choose the best prediction mode. For this, the proposed SS method defines the following two additional prediction modes:

- **SS mode** – In this case, SS estimation is used to find a predictor block (as explained in Section 4.1.1.1) for encoding the current block. As in HEVC inter coding, the eight partition patterns, i.e., M×M, M×(M/2), (M/2)×M, (M/2)×(M/2), M×(M/4), M×(3M/4), (M/4)×M, and (3M/4)×M (as shown in Figure 3.5), are allowed to define a flexible way to partition the CB for the SS estimation process in Figure 4.3b. In order to further improve the RD performance of this process, a jointly estimated Bi-SS prediction is proposed in Section 4.4. Moreover, the search window **W** (see Figure 4.3b) is restricted to minimize the encoding computational complexity.
- **SS-skip mode** – The SS-skip is employed only for the M×M partition pattern, and, in this case, the SS vector is directly derived from the HEVC merge technique [123] (seen in Section 3.2.3.2) and no further information is transmitted. However, the SS-skip mode restricts the SS vector candidates from only spatially neighboring CBs and guarantees that the area referenced by the chosen SS vector is already available in the SS reference at decoding time. Improvements to this method for the SS vector prediction are proposed in Section 4.3.

4.1.1.3 SS compensation

For each CB, the SS reference is updated by including the inverse quantized and inverse transformed causal prediction residual added to the predictor block to obtain the reconstructed samples as it will be available at the decoder side.

4.1.1.4 SS Vector Prediction

As discussed in Chapter 3 (see Section 3.2.3.1), motion vectors are predictively coded in HEVC by using the tool called AMVP. In this case, a vector candidate list is constructed by selecting vectors from CBs in the spatial and temporal (co-located) neighborhood.

Regarding SS vector prediction, the AMVP is also used but limiting the AMVP list to only vector candidates from spatially neighbor CBs. To further improve the RD performance of this scheme for LF image coding, improvements to the AMVP are also proposed in Section 4.3.

4.1.2 Reference Picture Management

To allow the SS estimation and SS compensation in intra-coded frames of HEVC, the reference lists construction and signaling need to be altered so as to include the SS reference. This process is similar to the temporal lists construction on HEVC inter-coded frames, and is managed by the general coder control block in Figure 4.2 by using the concept of RPS (see

Section 3.2.6). For this, the SS reference is made available to be used as a reference at the decoded picture buffer (see Figure 4.2).

4.1.3 LFC Uni-SS versus HEVC Still Picture Profile

This section illustrates some preliminary experimental results of the proposed LFC solution that motivated the proposals for further improvements described in Sections 4.3 and 4.4.

For this, the RD performance of the proposed LFC solution using Uni-SS prediction is here presented for LF image coding and compared against HEVC Main Still Picture Profile [2]. In these tests, the HEVC reference software version 14.0 [209] is used as the benchmark, as well as the base software for implementing the proposed codec.

For this, eight different LF images are considered: i) *Fredo*, as shown in Figure A.1; ii) *Seagull* (Figure A.4); iii) *Laura* (Figure A.3); iv) *Jeff* (Figure A.2); v) *Zhengyun1* (Figure A.5); vi) *Demichelis Spark* (first frame of the LF video sequence shown in Figure A.7); vii) *Plane and Toy* (frame number 123 of the LF video sequence shown in Figure A.8b); and viii) *Robot 3D* (frame number 54 of the LF video sequence shown in Figure A.9). RD performance is evaluated here through the Bjøntegaard Delta (BD) [210] metrics , i.e., in terms of the luma PSNR of the LF image and the corresponding Bitrate (BR) in terms of Bits Per Pixel (bpp).

From these preliminary results (Table 4.1), it can be observed that introducing the Uni-SS prediction into HEVC is advantageous for exploiting the inherent LF image correlations, leading to BD gains of up to 3.47 dB with 45.94 % of bit savings compared to HEVC Main Still Picture Profile [2].

Table 4.1 RD performance (using BD metrics [210]*) of LFC Uni-SS versus HEVC Main Still Picture Profile (QP values 27, 32, 37, and 42)*

LF Image	**LFC Uni-SS vs. HEVC Still Picture Profile**	
	PSNR [dB]	**BR [%]**
Fredo	**3.47**	**-45.94**
Seagull	3.33	-52.03
Laura	2.74	-41.26
Jeff	2.95	-45.33
Zhengyun1	2.48	-44.27
Demichelis Spark (frame 1)	2.19	-42.89
Plane and Toy (frame 123)	1.90	-25.10
Robot3D (frame 54)	1.22	-15.83
Average	**2.54**	**-39.08**

4.2 Extending LFC Uni-SS for LF Video Coding

Since an LF video is, actually, a sequence of 2D frames, the LFC Uni-SS solution can be straightforwardly extended for LF video coding by enabling Uni-SS prediction in conventional HEVC inter-coded frames. In this sense, this section briefly illustrates that, although the proposed LFC solution depicted in Figure 4.2 targets still LF image compression, it is also possible and advantageous to extend the solution for LF video coding.

For this, since there is only one SS reference picture for each picture to be coded, no additional signaling is necessary to distinguish between an SS reference and a temporal reference picture, and the distinction can still be based on the POC distance to the current picture (see Section 3.2.6). Furthermore, extensive tests were carried out to determine the most efficient position in which to include the SS reference in the list of reference pictures. It was concluded that the list should be firstly initialized with the short-term temporal reference pictures that are indicated in the RPS (see Section 3.2.6). Hence, the SS reference is appended to the list afterwards, becoming available for prediction of the current picture. Notice that, if the number of active reference pictures is kept the same (i.e., if it is not increased by one to include the SS reference), the computational complexity load will be identical to that of conventional HEVC inter-coded frames.

To assess the performance of the LFC Uni-SS solution for inter-coded frames, four different LF test sequences are considered: i) *Demichelis Cut*; ii) *Demichelis Spark;* iii) *Plane and Toy*; and iv) *Robot 3D*. A detailed description about the characteristics of these LF sequences is shown, respectively, in Tables A.2 to A.5. For these tests, the HEVC reference software version 14.0 is used as the benchmark, as well as the base software for implementing the proposed LFC Uni-SS Inter codec. In both solutions (i.e., LFC Uni-SS Inter and HEVC Inter), the abovementioned test sequences are encoded according to Low-delay P configuration [176].

The results are shown in Table 4.2 in terms of BD-PSNR and BD-BR [210] against HEVC Inter for the QP values 27, 32, 37, and 42. From these results, it can be seen that the LFC Uni-SS Inter solution also outperforms HEVC for inter-coded frames with gains of up to 1.08 dB and -34.30 % of bit savings.

However, it is shown in [59] that the RDO used in HEVC is not appropriate for LF video coding using the proposed LFC solution, since it is not able to find the best balance between rate and distortion constraints. Moreover, it is shown that further RD gains can be achieved when the RDO Lagrangian multiplier (see (3.1)) is better adjusted for intra- and inter-coded frames.

Table 4.2 RD performance (using BD metrics [210]*) of LFC Uni-SS Inter versus HEVC Inter video coding for selected LF video test sequences (QPs 27, 32, 37, and 42)*

LF Sequence	**LFC Uni-SS Inter vs. HEVC Inter**	
	PSNR [dB]	**BR [%]**
Demichelis Cut	0.86 dB	-27.13 %
Demichelis Spark	**1.08 dB**	**-34.30 %**
Plane and Toy	0.53 dB	-11.29 %
Robot 3D	0.43 dB	-8.22 %
Average	**0.73 dB**	**-20.24 %**

Therefore, investigating more suitable RD models and rate control methods (to replace these used in HEVC) for LF video applications will be left as future work.

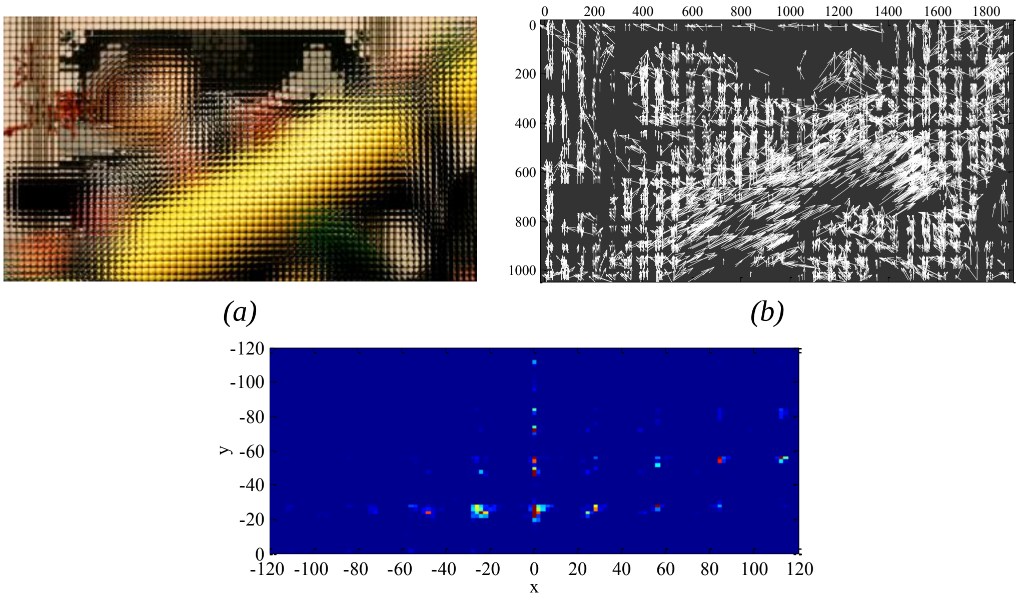
4.3 Proposal for an Improved SS Vector Prediction

Although the Uni-SS prediction is able to significantly improve the HEVC RD performance for LF image coding, as illustrated in Section 4.1.3, it was observed that the conventional AMVP and merge mode (see Section 3.2.3) were not able to take advantage of the distinctive characteristics of the SS prediction data. Specifically, due to the following reasons:

1) The definition of SS co-located CB positions is not applicable since the SS reference is the current frame itself. For this reason, in intra-coded frames, AMVP and merge candidate lists are limited to spatial candidates only. Furthermore, zero vector candidates are never available for a SS reference – as it corresponds to a position outside the causal area **W** (see Figure 4.3b) – making the number of candidate vectors even more limited.

2) Besides the possible correlation between vectors from neighboring CBs, the SS vectors present a regular distribution, which is related to the structure of the MIs. As an example, Figure 4.4b shows the distribution of the transmitted SS vectors for an LF image with MI resolution of 28×28, and Figure 4.4c shows a heat map of the corresponding SS vector values distribution. From this, it can be observed that the magnitude of the SS vectors is usually multiple of the MI resolution so as to compensate the micro-image boundaries.

Therefore, this section proposes an improved SS vector prediction for enhancing the RD performance of the proposed LFC solution. Moreover, it is shown that further bit savings can be achieved for the LFC solution using Uni-SS prediction, with respect to the experimental results presented in Section 4.1.3.



(a)

(b)

(c)

Figure 4.4 Characteristics of the SS prediction data in a coded LF image: (a) The original LF image with MI resolution of 28×28; (b) SS vector distribution along the LF image; and (c) Heat map showing the corresponding vector values distribution

4.3.1 The MIVP Vector Candidates

Based on the specific nature of SS vectors (see Figure 4.4), a set of new candidate SS vectors, referred to as Micro-Image-Based Vector Predictors (MIVP), is proposed to be included into AMVP and merge mode of HEVC to further improve the proposed LFC coding solution.

This new set includes the left MIVP vector, above MIVP vector, and above-left MIVP vector, as given in (4.2), (4.3), and (4.4), respectively. Where $PB_h \times PB_v$ corresponds to the current PB size, and $MI_j \times MI_i$ represents the MI resolution (in pixels).

The terms $\lceil PB_h/MI_j \rceil$ and $\lceil PB_v/MI_i \rceil$ force the candidate vectors to be distributed according to the structure of MIs**,** as well as to be inside the area of previous coded and reconstructed CBs. Figure 4.5 shows that the position of the three vector candidates is always related to the MI structure independently of the current PB partition size (that can be larger or smaller than the MI resolution).

$$L_{MIVP_{vector}} = \left(-\left\lceil \frac{PB_h}{MI_j} \right\rceil \times MI_j \text{ },0 \right) \tag{4.2}$$

$$A_{MIVP_{vector}} = \left(0\text{ }, -\left\lceil \frac{PB_v}{MI_i} \right\rceil \times MI_i \right) \tag{4.3}$$

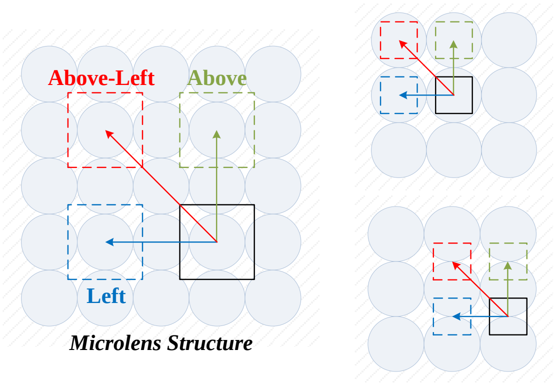


Figure 4.5 Proposed MIVP candidates. Examples when the PB is larger (left) and smaller (right) than the MI resolution

$$AL_{MIVP_{vector}} = \left(-\left\lceil \frac{PB_h}{MI_j} \right\rceil \times MI_j \,, -\left\lceil \frac{PB_v}{MI_i} \right\rceil \times MI_i \right) \tag{4.4}$$

Finally, to introduce the proposed MIVP candidates into AMVP and merge mode, the following changes (regarding the conventional HEVC AMVP and merge mode) are needed:

- **Allowing MIVP Candidate in AMVP** – for the SS reference, after the selection of conventional AMVP vector candidates, one MIVP candidate is selected between the left, above, or above-left MIVP vector (in this order). In the final AMVP list there are up to three candidates (as in [2]) between spatial and MIVP candidates.

- **Allowing MIVP Candidates in Merge** – after selecting the available conventional merge candidates, up to three MIVP merge candidates (i.e., MIVP vectors plus the position of the SS reference in the list of reference pictures) are included into the merge candidate list until the maximum number of candidates is reached. The order of selection is defined as left, above and above-left. As in [2], the maximum number of merge candidates is five.

Notice that the used MI resolution, $MI_j \times MI_i$, can be an approximate value that may be easily derived from the texture information. Moreover, this information can be derived at the encoder side and transmitted once in the high level syntax elements of the bitstream, resulting in nearly no performance loss.

4.3.2 MIVP Efficiency for Uni-SS Prediction

To illustrate the advantage of adding MIVP candidates into the Uni-SS prediction, the results presented in Section 4.1.3 for the Uni-SS (without MIPV) are here compared to the Uni-SS with MIVP enabled. For this, the corresponding BD [210] results are shown in Table 4.3 using the same test conditions that were adopted in Section 4.1.3.

Table 4.3 RD performance (using BD metrics [210]*) of LFC Uni-SS with MIVP*

LF Image	Uni-SS w/ MIVP vs. Uni-SS w/o MIVP	
	PSNR [dB]	**BR [%]**
Fredo	0.37	-7.21
Seagull	0.27	-6.63
Laura	0.11	-2.48
Jeff	0.21	-4.67
Zhengyun1	0.20	-5.17
Demichelis Spark (frame 1)	**0.30**	**-8.29**
Plane and Toy (frame 123)	0.20	-3.28
Robot3D (frame 54)	0.13	-1.91
Average	**0.22**	**-4.95**

From these results (Table 4.3), it can be seen that by using the proposed MIVP it is possible to improve the Uni-SS prediction RD performance, leading to further bitrate savings of up to 8.29 % and coding gains of up to 0.30 dB, when compared to LFC Uni-SS without MIVP.

4.4 Proposal for a Jointly Estimated Bi-SS Prediction

To further improve the performance of the proposed LFC solution, a novel jointly estimated Bi-SS estimation and compensation scheme, which is based on the generic concept of superimposed prediction, is here proposed to replace the aforementioned Uni-SS estimation and compensation processes.

To motivate the adoption of this Bi-SS prediction in the LFC solution presented in Section 4.1, a theoretical analysis is firstly presented, which shows that other spatial displacement compensated prediction schemes – such as the IntraBC [193–195], the preceding Uni-SS prediction proposed in Section 4.1.1.1, and the solution for bi-prediction proposed in [189] – can be considered as restricted cases of the Bi-SS solution proposed here. Additionally, the influence of the MIVP in the RD performance achieved by the Bi-SS prediction is also presented.

4.4.1 Theoretical Bi-SS Performance Analysis

The RD performance improvement due to the adoption of the jointly estimated Bi-SS prediction (presented in more details in the following section) is based on three main hypotheses, which will be analyzed in this section:

1) With a large enough search window, W, (see Figure 4.3b), it is possible to find two predictor blocks that properly represent the current block, I (x), i.e., with low residual signal.

2) By combining two good predictor blocks, it is possible to further minimize the residual signal of the SS compensated prediction, compared to only using the uni-predicted SS candidate (as in the Uni-SS prediction presented in Section 4.1.1.1 and in the IntraBC scheme [193–195]).

3) Jointly estimating the predictor blocks leads to better RD performance than deriving them independently (as in reference software for HEVC inter B frame coding [209] and in the bi-prediction solution proposed in [189]).

For this analysis, the performance of the proposed Bi-SS prediction is here modeled by the uncertainty [52] (or inaccuracy [51, 211]) in the SS compensated prediction signal.

Regarding the first abovementioned hypothesis, it is valid due to the following facts:

- Given the small baseline between adjacent microlenses in the acquisition process, a significant cross-correlation exists between neighboring MIs, as shown by the autocorrelation function in Figure 4.3a. It can be seen that the autocorrelation function presents a regular structure of spikes and the constant distance between these regular spikes corresponds to the MI spacing in the array [59]. Since these highly correlated samples are distributed along the MIs, it is likely that similarly good predictor blocks will be also distributed accordingly.

- It was shown in [59] that, when using the SS compensated prediction scheme for exploiting the inherent MI cross-correlation, the distribution of the chosen SS vectors is also related to the size and arrangement of the MIs in the array. This can be illustrated by the heat map in Figure 4.4c, where brighter areas correspond to more frequent SS vector amplitudes. Hence, since these most frequent best uni-predicted SS vectors are distributed in all directions according to the MI arrangement, it is possible to consider that the second best SS vector, which can also represent the current block properly, is likely to be found according to this distribution in a different direction.

Regarding the second and third hypothesis, the residual signal for the Bi-SS compensated prediction is given by (4.5),

$$e(\mathbf{x}) = I(\mathbf{x}) - \sum_{p=0}^{1} h_p(\mathbf{x}) \cdot \tilde{I}(\mathbf{x} - \mathbf{v}_p) \tag{4.5}$$

where $\mathbf{x} - \mathbf{v}_\mathrm{p} \in \mathbf{W}$, and h_p is the weight for each of these predictor blocks. For instance, for the Bi-SS prediction proposed in Section 4.4.2, $h_p = 1/2$, $\forall p \in \{0,1\}$. However, as discussed in [52], the residual signal given by (4.5) can be actually generalized as in (4.6) for N predictor blocks.

$$e(\mathbf{x}) = I(\mathbf{x}) - \sum_{p=0}^{N-1} h_p(\mathbf{x}) \cdot \tilde{I}(\mathbf{x} - \mathbf{v}_p) \tag{4.6}$$

The general case represented by (4.6) can also incorporate other types of candidate predictors reflecting the very flexible set of inter coding tools of HEVC [2]. In this case, the $h = (h_0, h_1, \ldots, h_{N\text{-}1})$ corresponds to a weight vector that is able to, for example, incorporate [52]: i) the filtering used to generate the quarter-pel interpolated signal in the SS estimation; and ii) the deblocking filter that can be applied in the SS reference. Each $\tilde{I}_p\,(\mathbf{x})=\tilde{I}\,(\mathbf{x}\text{-}\mathbf{v}_p)$ term can be interpreted as each of the multiple compensated signals available for prediction of the current block. Hence, the uncertainty in a given Bi-SS compensated prediction can be modeled, as in [52], by an a posteriori probability density function, $h_p(\mathbf{x})$, conditioned on the encoded data. Therefore, since the expected value (the second term on the right-hand side of (4.6)) is the estimator that minimizes the mean-square error in the prediction of a random variable [52], it is possible to say that the residual signal in (4.6) can be minimized and, consequently, the accuracy of the prediction can be improved by using a larger set of multiple compensated signals and an optimized weight vector [52].

In addition, another possibility is to analyze the performance of the jointly estimated Bi-SS prediction by modeling the inaccuracy of each used displacement vector $\mathbf{v}_p$, as in [51, 211]. For this, Figure 4.3b shows that, although the pixel correlation in the LF image is not as smooth as in conventional 2D images, each MI itself has some degree of inter-pixel redundancy as in common 2D images (see Figure 4.3b). Thus, it is possible to consider that samples inside each MI follow the same correlation model as samples in a 2D image (i.e., an isotropic exponentially decaying autocorrelation function). This assumption is reasonable at least for $I\,(\mathbf{x})$ smaller than the MI resolution, and has been also adopted in [189, 212] for LF images. With this assumption, the accuracy of the SS compensation can be measured by the displacement error variance [211], and the same signal model used in [51, 211] can also be considered for the SS compensated prediction signal. In this case [51, 211], **h** denotes a row vector of impulse responses $(h_0, h_1, \ldots, h_{N\text{-}1})$ of a 2D prediction filter [211], and the residual signal is given by (4.7),

$$e(\mathbf{x}) = I(\mathbf{x}) - \mathbf{h}(\mathbf{x}) * \tilde{\mathbf{I}}(\mathbf{x}) \tag{4.7}$$

where the second term on the right-hand side of the equation denotes a 2D convolution of the prediction filter **h** to a column vector of N multiple compensated signals $\tilde{\mathbf{I}} = (\tilde{I}_0, \tilde{I}_1, \ldots, \tilde{I}_{N\text{-}1})^{\mathrm{T}}$. In this model, both I and each component of $\tilde{\mathbf{I}}$ are assumed to be wide sense stationary random processes with an additive Gaussian noise signal. Hence, these noisy signals may comprise all signal components of the SS compensated prediction that cannot be described by the translational displacement model [211].

Based on the abovementioned approximations, the conclusions from [51, 211] also hold for validating the second and third hypothesis. Notably, with high rate assumptions:

- Concerning the second hypothesis, the optimal filter **h** (i.e., that minimizes the mean square error) can be interpreted as a low pass filter that removes high frequency components from $\tilde{\mathbf{I}}$ that are too noisy or that change too rapidly [211]. From the theoretical analysis in [211], it was concluded that increasing the number of equally good predictor blocks always led to bitrate savings compared to a more limited set of predictor blocks, even if the simple average filter is used, instead of considering an optimal filter $\mathbf{h}$ in (4.7). Therefore, this suggests that increasing from one predictor block (as in the Uni-SS prediction) to two predictor blocks (as in the Bi-SS solution proposed in this section) minimizes the residual signal in the SS compensated prediction.

- Concerning the third hypothesis, an extended analysis was performed in [51] for the case where the multiple compensated samples of $\tilde{\mathbf{I}}$ are jointly estimated. In this case, the displacement error of all components of $\tilde{\mathbf{I}}$ are assumed to be correlated, instead of being independent as assumed in [211]. Moreover, this analysis considered the simple average filtering case as for the proposed Bi-SS. Therefore, it was shown that a combination of two jointly estimated predictor blocks is more efficient than two independent predictor blocks [51]. Furthermore, it was concluded that, for jointly estimated predictors, the major portion of the gain is already achievable by only two predictor blocks [51]. This suggests that further rate-distortion gains can be achieved by jointly estimating the two predictor blocks for bi-prediction compared to simply combining two best uni-predicted candidates (as in the HEVC reference software inter B frame coding [2] and in bi-predicted solution proposed in [189]).

Therefore, without loss of generality, the IntraBC scheme in [193–195], the Uni-SS prediction presented in Section 4.1.1.1, and the solution proposed in [189] can be seen as restricted cases of the Bi-SS solution being proposed in this section. In these cases, restrictions are imposed in the number of predictor blocks, the number of allowed partition patterns and sizes, and in the bi-prediction estimation process for each predictor block that is independently employed in different areas of the SS reference.

As it will be seen in Section 4.5, the theoretical insights from this section for the proposed Bi-SS solution are supported by the experimental results for LF images. Nevertheless, further prediction performance improvements are still possible for the LFC Bi-SS solution, notably, by using an optimized filter $\mathbf{h}$ for bi-prediction in (4.7). For this reason, future work will include studying proper algorithms for estimating an optimal filter $\mathbf{h}$.

4.4.2 Bi-SS Candidate Predictor Estimation

Motivated by the theoretical analysis presented above, the proposed Bi-SS prediction is here presented, which is based on the generic concept of superimposed prediction [51].

More specifically, there is only a single reference picture available in the Bi-SS compensated prediction, i.e., the SS reference [60, 61], and only two possible candidate predictors (instead

```
Initialization:   $k = 0,\quad \mathbf{v}_0^{(k)} = \mathbf{v}_0^{best},\quad J = J_{MAX}$
 do
    $J_{best} = J$
    $p = (k) \bmod (2)\,,\quad q = 1 - p$
    $J = \left\| I(\mathbf{x}) - \frac{\tilde{I}\left(\mathbf{x}-\mathbf{v}_q^{(k+1)}\right) + \tilde{I}\left(\mathbf{x}-\mathbf{v}_p^{(k)}\right)}{2} \right\|_1 + \lambda \left[ R\left(\mathbf{v}_q^{(k+1)}\right) + R(\mathbf{v}_p^{(k)}) \right]$
    $\mathbf{v}_q^{(k+1)} = \underset{\mathbf{v}_q^{(k+1)}}{\operatorname{argmin}}\ J$
    $k = k+1$
 while $k \neq K$ or $J < J_{best}$
```

Figure 4.6 Algorithm for jointly estimating the two predictor blocks $\tilde{I}(\mathbf{x}\text{-}\mathbf{v}_0)$ and $\tilde{I}(\mathbf{x}\text{-}\mathbf{v}_1)$ for the proposed Bi-SS candidate predictor. The index q defines which of the two vectors ($\mathbf{v}_0$ or $\mathbf{v}_1$) will be optimized in a particular iteration k, while the index p defines the vector that will be kept fixed

of the three candidates of HEVC [2] that are used in [189]) are derived to predict the current block, namely: i) the uni-predicted SS candidate; and ii) the bi-predicted SS candidate.

The uni-predicted SS candidate corresponds to the previously proposed Uni-SS prediction (see Section 4.1.1.1), in which the predictor block is found by minimizing the Lagrangian cost function in (4.1).

The proposed bi-predicted SS candidate predictor differs from the HEVC reference software inter B frame bi-prediction, as well as from the bi-predicted solution in [189], for two main reasons:

1) The two predictor blocks in the Bi-SS solution are derived from the same reference picture (the SS reference) and no additional signaling is needed for reference indices. More specifically, these two predictor blocks are estimated in the same search window, **W** (see Figure 4.3b), and, consequently, they can be located in the same MI and in overlapped pixel positions (as illustrated in the dashed blue lines in Figure 4.3b).
2) To further improve the prediction efficiency, these two predictor blocks are jointly estimated in the complete search window (as explained below), instead of combining two best uni-predicted candidates (as in [189]).

For jointly estimating the two predictor blocks, the locally optimal rate-constrained algorithm proposed in [58] is used (see Figure 4.6). This algorithm avoids searching through all possible combinations of two candidate blocks $\tilde{I}(\mathbf{x}\text{-}\mathbf{v}_0)$ and $\tilde{I}(\mathbf{x}\text{-}\mathbf{v}_1)$ inside **W**. For this, in each algorithm iteration, k, an optimal SS candidate vector $\mathbf{v}_q^{(k+1)}$ (with index $q \in \{0,1\}$) is found by minimizing the Lagrangian cost function conditioned to the optimal SS candidate vector found in the previous iteration $\mathbf{v}_p^{(k)}$ (with $p \in \{0,1\}$). Therefore, the algorithm is focused on finding an optimized vector $\mathbf{v}_1$ conditioned to a known vector $\mathbf{v}_0$ in even iterations, and vice versa in odd iterations. For instance, in the first iteration, $k = 0$, $p = 0$, $q = 1$, and the optimal $\mathbf{v}_1^{(1)}$ is found by fixing $\mathbf{v}_0^{(0)}=\mathbf{v}_0^{best}$ (the best uni-predicted SS candidate vector found in (4.1)). Similarly, in the second iteration, $k = 1$, $p = 1$, $q = 0$, and the optimal $\mathbf{v}_0^{(2)}$ is found by fixing $\mathbf{v}_1^{(1)}$ (which was found in the previous iteration).

The maximum number of iterations, K, defines a tradeoff between complexity and RD performance and can be adjusted according to the system constraints. In this Thesis, $K = 2$ and, consequently, the corresponding complexity is similar to that of HEVC inter B-frame with one active reference in each reference picture list.

Finally, the best candidate among uni-predicted SS and bi-predicted SS is chosen by conventional RDO [52].

4.4.3 Bi-SS Vector Prediction

If bi-prediction is used, two predictor vectors are derived from the AMVP method (one for each SS estimated vector) and the difference between the two SS estimated vectors and the corresponding predictor vectors is transmitted along with the indices of the chosen candidates in the list. In the case of the SS-skip mode, bi-predicted merge candidates may be derived from neighboring CBs that were coded with Bi-SS mode. Furthermore, if the merge candidate list is not fully populated, bi-predicted candidates can also be derived by combining two existing candidates from different reference picture lists.

In addition to this, the set of new MIVP candidate vectors, presented in Section 4.3, is also included into AMVP and merge candidate lists to further improve the RD performance. For this reason, the following section briefly analyzes the efficiency of the MIVP candidate vectors for Bi-SS prediction.

4.4.4 MIVP Efficiency for Bi-SS Prediction

To illustrate the RD efficiency of using the MIVP vector prediction also for Bi-SS prediction, the performance of Bi-SS prediction with MIVP (referred to as Bi-SS w/ MIVP) and without MIVP (referred to as Bi-SS w/o MIVP) is here compared.

For this, the BD [210] results using the same test conditions adopted in Section 4.1.3 are presented in Table 4.4. These results can be also compared to the MIVP RD performance for Uni-SS prediction, which can be found in Table 4.3.

From these results, it is possible to see that the gains of including the MIVP candidate vectors for Bi-SS prediction (Table 4.4) are slightly lower than for Uni-SS prediction (Table 4.3). However, the MIVP is still relevant for the LFC Bi-SS solution, leading to further bit savings of up to 6.58 %.

Table 4.4 BD performance (using BD metrics [210]*) of the MIVP for the proposed Bi-SS solution,*

LF Image	**LFC Bi-SS w/ MIVP vs. LFC Bi-SS w/o MIVP**	
	PSNR [dB]	**BR [%]**
Fredo	0.35	**-6.58**
Seagull	0.23	-5.52
Laura	0.11	-2.32
Jeff	0.19	-4.22
Zhengyun1	0.18	-4.65
Demichelis Spark (frame 1)	0.12	-3.47
Plane and Toy (frame 123)	0.19	-3.08
Robot3D (frame 54)	0.10	-1.55
Average	**0.18**	**-3.92**

4.5 Performance Evaluation of the Complete LFC Bi-SS Solution

This section assesses the performance of the proposed LFC codec (see Figure 4.2), which comprises the MIVP vector prediction proposed in Section 4.3 as well as the jointly estimated Bi-SS prediction proposed in Section 4.4. This complete solution is here simply referred to as LFC Bi-SS.

For this, the evaluation methodology is firstly introduced and, then, the obtained results are presented and discussed.

4.5.1 Evaluation Methodology

The methodology to evaluate the performance of the proposed LFC Bi-SS solution can be summarized in the following sub-sections.

4.5.1.1 Test Images

Eight LF test images with different optical acquisition setups and scene characteristics are used (see Appendix A) so as to achieve representative RD results. These are: (i) *Fredo*, as shown in Figure A.1; (ii) *Seagull* (Figure A.4); (iii) *Laura* (Figure A.3); (iv) *Jeff* (Figure A.2); (v) *Zhengyun1* (Figure A.5); (vi) *Demichelis Spark* (first frame of the sequence shown in Figure A.7); (vii) *Plane and Toy* (frame number 123 of the sequence shown in Figure A.8b); and (viii) *Robot 3D* (frame number 54 of the sequence shown in Figure A.9).

The complete characterization of these LF images is shown in Tables A.1-A.5. These images were converted to Y'CbCr 4:2:0 color format and pre-processed (see Appendix A) before being encoded. It is worthy stressing here that the pre-processing presented in Appendix A is not actually mandatory for the proposed LFC coding solution. However, it is adopted for the experimental results presented in this section in order to fairly comparing the proposed LFC solution to other tested solutions in which this pre-processing is essential, as well as to make the results homogeneous across all chapters.

4.5.1.2 Test Conditions

The experimental results consider the following test conditions:

- **Codec Software Implementation** – The reference software of HEVC version 14.0 [209] is used as the benchmark, as well as the base software for implementing the proposed codec.
- **Search Range** – A search range value of 128 is adopted for all tested LF images (i.e., w=128 in Figure 4.3b).
- **Search Strategy** – The full search algorithm with the HEVC quarter-pixel accuracy is used.
- **Coding Configuration** – The results are presented for five QP values (22, 27, 32, 37, and 42).
- **RD Evaluation** – The BD [210] results are presented in terms of the luma PSNR of the raw LF image and the corresponding rate in terms of bpp values. When analyzing the results in terms of the BD-PSNR and BD-BR, the sets of QP values {22, 27, 32, 37} and {27, 32, 37, 42} are considered for analyzing the performance, respectively, for high and low bpp values.
- **Rendering-Dependent Objective Quality Metrics** – In addition to the luma PSNR of the entire LF image, two other objective quality metrics are used in order to provide a better correspondence to the quality perception expected by a user positioned in front of a future LF display or the quality of 2D views synthesized from the coded and reconstructed LF image. In this context, two rendering-dependent objective metrics are here adopted, namely: i) Rendering-dependent PSNR, and ii) Rendering-dependent SSIM. Similarly to the metric proposed in [93], the average (and also the standard deviation) PSNR and SSIM are calculated from a set of K views rendered from the LF image. These metrics are computed, respectively, through (4.8) and (4.9), considering only the luma component. Note that any rendering algorithm may be used; however, each of them may differently affect the quality of the rendered views, which may make difficult the coding performance evaluation alone (for this reason these metrics are here name as 'Rendering-dependent'). For the results presented in Section 4.5.2, instead of adopting the algorithm proposed in [93], the algorithm referred to as Basic Rendering [10] (see Chapter 2) is used to generate the views (since the algorithm proposed in [93] requires exact information of the used acquisition setup, which is not available for some of the LF test images).

$$\begin{aligned}\overline{\text{PSNR}}_{\text{Rendering-dependent}} &= \frac{1}{K}\sum_{k=1}^{K} PSNR_k \\ \sigma_{\text{Rendering-dependent}} &= \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(PSNR_k - \overline{\text{PSNR}}\right)^2}\end{aligned} \tag{4.8}$$

$$\begin{aligned}\overline{\text{SSIM}}_{\text{Rendering-dependent}} &= \frac{1}{K}\sum_{k=1}^{K} SSIM_k \\ \sigma_{\text{Rendering-dependent}} &= \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(SSIM_k - \overline{\text{SSIM}}\right)^2}\end{aligned} \tag{4.9}$$

- **Visual Quality Evaluation** – Additionally, a portion of a central view rendered from the coded and reconstructed LF image is used for a visual inspection in Section 4.5.3. For rendering the views, the Basic Rendering algorithm is also used. For all compared coding solutions, the QP values of the encoder are adjusted to lead to the same bpp values for all LF images. Notice that, since there is still no consensus on the scientific community regarding subjective evaluation methodologies for LF content, the results shown Section 4.5.3 are presented as an illustrative qualitative analysis of the proposed LFC Bi-SS coding solution.

4.5.1.3 Benchmark Solutions

In order to assess the performance of the proposed coding architecture (see Figure 4.2), the LFC Bi-SS solution is compared to solutions where the LF image is decomposed into a PVS (or multiview) content. For this, the following four coding solutions are tested, similarly to what has been proposed in [92, 175]:

1) **MI-Based PVS (Low Delay P)** – A PVS of MIs is coded using HEVC. Since the resolution of each MI is considerably smaller than HEVC commonly supported resolutions, the largest CB size was set to 16×16. The MI-based PVS is then encoded using the Low Delay P [176] configuration (see Figure 3.15d). However, in order to fairly use the metrics in (4.8) and (4.9), the QP values are kept the same for all MIs in the PVS. Various MI scanning orders were tested (i.e., raster, parallel, zig-zag, and spiral, as illustrated in Figure 3.11), but only the spiral order is presented as it achieved the best RD performance.

2) **MI-Based PVS (Random Access)** – The PVS of MIs scanned in spiral order is encoded using HEVC using Random Access [176] configuration (see Figure 3.15d). Similar to the previous solution, i.e., MI-based PVS (Low Delay P), the largest CB size is set to 16×16 and the QP values are kept the same for all MIs in the PVS.

3) **VI-Based PVS (Low Delay P)** – In this case, a PVS of VIs is encoded with HEVC. After testing various orders for scanning the VIs, the spiral order is presented as it

achieved the best RD performance. The PVS of VIs is then encoded using the same conditions as defined for the MI-based PVS (Low Delay P) solution.

4) **VI-Based PVS (Random Access)** – The PVS of VIs scanned in spiral order is encoded with HEVC using the same coding conditions defined for MI-based PVS (Random Access).

Additionally, to assess the RD performance for different local and non-local spatial prediction schemes, five HEVC-based coding solutions are also compared to the proposed LFC Bi-SS. To guarantee a fair comparison between all of them, the same test conditions, shown in Section 4.5.1.2 are also adopted. These five benchmark solutions are:

1) **HEVC** – In this case, the LF image is encoded with HEVC, using the Main Still Picture profile [2].

2) **HEVC RExt** – In this case, the LF image is encoded using the HEVC RExt reference software version 6.0 [193], where IntraBC prediction is used. As discussed in this chapter, this solution is a restricted case of the proposed LFC Bi-SS since a reduced set of coding options is used. Specifically, this is due to the usage of: i) limited partition patterns (i.e., $M\times M$, $M\times(M/2)$, $(M/2)\times M$, and $(M/2)\times(M/2)$); ii) limited CB sizes (CBs larger than 16×16 are skipped, based on a threshold on RD cost); iii) only 1D vectors for 16×16 CBs; iv) integer block matching search across only one or more CBs to the left; and v) a default vector predictor instead of AMVP (notably, the vector of the latest coded CB).

3) **HEVC SCC** – The original LF image is encoded using the SCC extension of HEVC, reference software SCC 1.0 [195]. It is worth noting that some improvements in performance for screen content coding were proposed in this solution compared to the HEVC RExt 6.0. Notably, the search window was expanded over the entire CB row or column (for 16×16 CBs), and over some positions in the entire picture by using a hash-based search (for 8×8 CBs).

4) **LFC Uni-SS** – In this case, the test images are encoded with the proposed LFC coded using only Uni-SS prediction, as proposed in Section 4.3. This means that, in this case, only the uni-predicted candidate is available for the SS estimation and compensation. However, this solution also uses the proposed MIVP candidate vectors for improving the coding performance (as in the proposed LFC Bi-SS).

5) **LFC Restricted-SS** – In this case, the test images are encoded with the author's implementation of the solution proposed in [189], where bi-prediction is also allowed by simply using the HEVC reference software inter B frame prediction. In this case, as explained in [189], the SS reference search area is separated into two different parts, which are assumed to be two different reference pictures [189]. Therefore, as in HEVC inter B-frame prediction, three candidate predictors can be derived: the two best uni-predicted candidates from each of the two reference pictures, and a linear combination

Table 4.5 LFC Bi-SS RD performance (using BD metrics [210] *for QP values 27, 32, 37, and 42) regarding PVS-based solutions for the LF images: (a) Fredo, (b) Seagull, (c) Laura, (d) Jeff, (e) Zhengyun1, (f) Demichelis Spark (frame 1), (g) Plane and Toy (frame 123), and (h) Robot 3D (frame 54)*

LF Image	**VI-based PVS (Low Delay P)**		**VI-based PVS (Random Access)**		**MI-based PVS (Low Delay P)**		**MI-based PVS (Random Access)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
(a)	9.18	-85.65	9.34	-84.17	2.56	-36.74	2.29	-33.85
(b)	5.92	-81.94	5.70	-80.55	2.32	-41.18	1.88	-34.77
(c)	6.58	-78.22	6.22	-75.19	1.64	-29.67	1.26	-23.65
(d)	7.32	-83.41	6.89	-80.68	2.11	-35.83	1.85	-32.54
(e)	7.45	-85.78	7.40	-83.94	2.26	-40.80	2.08	-38.73
(f)	6.87	-84.59	5.89	-77.39	4.55	-65.36	4.14	-63.42
(g)	6.82	-69.42	4.99	-57.49	3.76	-42.35	3.26	-38.82
(h)	7.87	-69.96	6.41	-63.60	3.37	-36.28	3.09	-34.81
Average	**7.25**	**-79.87**	**6.60**	**-75.38**	**2.82**	**-41.03**	**2.48**	**-37.57**

of them for bi-prediction. As discussed in Section 4.4.1, this solution can be seen as a restricted case of the Bi-SS prediction. It is also worthwhile to notice that, the solution presented in [189] does not include the MIVP candidate vectors that are used in both LFC Uni-SS and LFC Bi-SS solutions.

4.5.2 LFC Bi-SS RD Performance Evaluation

Tables 4.5 and 4.6 present the RD performance using the BD metrics [210] for 'low bpp' values of the proposed LFC Bi-SS against the benchmark solutions presented in the previous section. In addition to this, Figures 4.7 to 4.14 presents the RD performance for each LF image in terms of the Rendering-dependent PSNR and SSIM metrics.

From these results, the following conclusions can be derived:

- **Regarding Different Data Arrangements** – Comparing the results from Table 4.5, it is possible to conclude that data arrangements based on MIs – i.e., MI-based PVS (Low Delay P) and MI-based PVS (Random Access) – are shown to be considerably more efficient than the corresponding VI-based PVS solutions with the same coding configuration. A careful analysis of these results shows that the multiplexing process from the captured MIs to VIs may result in images with very low resolution and with significant aliasing artifacts [86], which are then difficult to predict and to compress. This was particularly observed for the LF images presented in Appendix A, which were captured using a focused LF camera setup. Further comparisons with respect to a larger set of LF optical acquisition setups will be considered in future work. For some more results considering Lytro Illum LF images, the reader can alternatively refer to [175].

Table 4.6 LFC Bi-SS RD performance (using BD metrics [210] *for QP values 27, 32, 37, and 42) regarding HEVC-based benchmark solutions for the LF images: (a) Fredo, (b) Seagull, (c) Laura, (d) Jeff, (e) Zhengyun1, (f) Demichelis Spark (frame 1), (g) Plane and Toy (frame 123), and (h) Robot 3D (frame 54)*

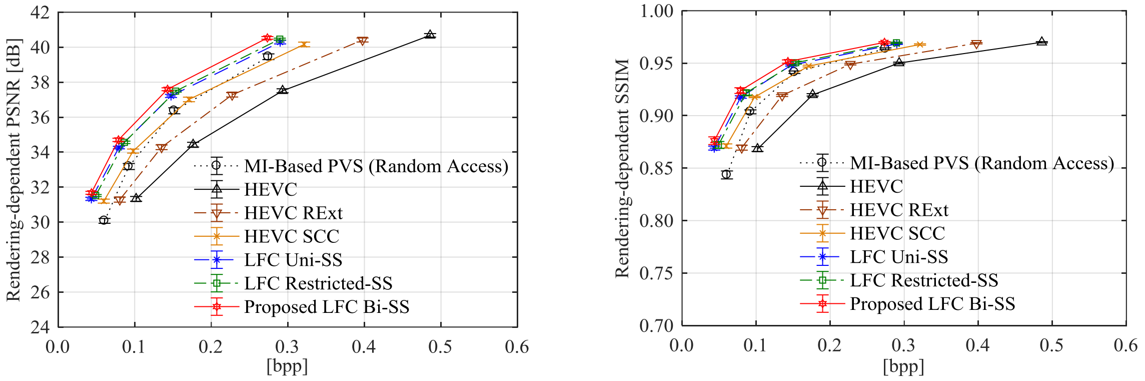
LF Image	**HEVC Still Pict. Profile**		**HEVC RExt**		**HEVC SCC**		**LFC Uni-SS**		**LFC Restricted-SS** [189]	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
(a)	4.23	-52.95	2.89	-41.12	1.46	-24.40	0.35	-6.80	**0.50**	**-9.32**
(b)	**4.29**	**-61.09**	**2.95**	**-48.60**	**1.77**	**-34.30**	0.58	-13.78	0.45	**-10.45**
(c)	3.35	-48.04	2.32	-37.67	1.21	-23.13	0.45	-9.46	0.21	-4.55
(d)	3.66	-52.81	2.60	-42.26	1.59	-29.41	0.45	-9.85	0.37	-8.00
(e)	3.11	-51.70	2.33	-42.77	1.37	-28.99	0.37	-9.06	0.37	-8.90
(f)	3.13	-55.96	2.01	-42.22	1.64	-36.98	0.65	**-17.32**	0.31	-8.76
(g)	2.42	-30.96	1.01	-14.91	0.81	-12.33	0.30	-4.89	0.40	-6.41
(h)	1.47	-18.87	0.42	-6.09	0.25	-3.74	0.12	-1.78	0.19	-2.77
Average	**3.21**	**-46.55**	**2.07**	**-34.45**	**1.26**	**-24.16**	**0.41**	**-9.12**	**0.35**	**-7.40**

Concerning the different prediction structures, it can be seen that the Random Access configuration performs better for both MI- and VI-based PVS data arrangements. Due to the above reasons, only the solution MI-based PVS (Random Access) is included for the results in Figures 4.7 to 4.14. Regarding the results in Figures 4.7 to 4.14, it can be seen that, although MI-based PVS (Random Access) outperforms HEVC for some LF images, it is always outperformed by the HEVC SCC solution.

- **Regarding Different HEVC-Based Coding Schemes** – As can be observed from Table 4.6 and Figures 4.7 to 4.14, the proposed LFC Bi-SS solution always outperforms the other HEVC-based benchmark solutions, presenting significant gains against HEVC (BD gains of up to 4.29 dB and 61.09 % of bit savings), HEVC RExt (up to 2.95 dB and 48.60 % of bit savings), and HEVC SCC (up to 1.77 dB and 34.30 % of bit savings). Regarding the LFC Uni-SS solution, it can be seen that increasing the number of predictor blocks from the LFC Uni-SS solution to the LFC Bi-SS solution leads to further bit savings (up to 17.32 %), as hypothesized in the theoretical analysis of Section 4.4.1. Moreover, comparing to the LFC Restricted-SS solution, it can be observed that further bit savings (up to 10.45 %) can also be achieved by jointly estimating both candidate blocks in the LFC Bi-SS solution, instead of devising them separately from different areas (as in the LFC Restricted-SS). In addition to these results, Appendix B presents some more results for coding Lytro Illum LF images using the proposed LFC Bi-SS solution (see Section B.2). It can be seen that similar conclusions can be derived, and the proposed LFC Bi-SS presents significantly better RD performance with gains (using BD metrics [210]) of up to 1.88 dB and 66.97 %

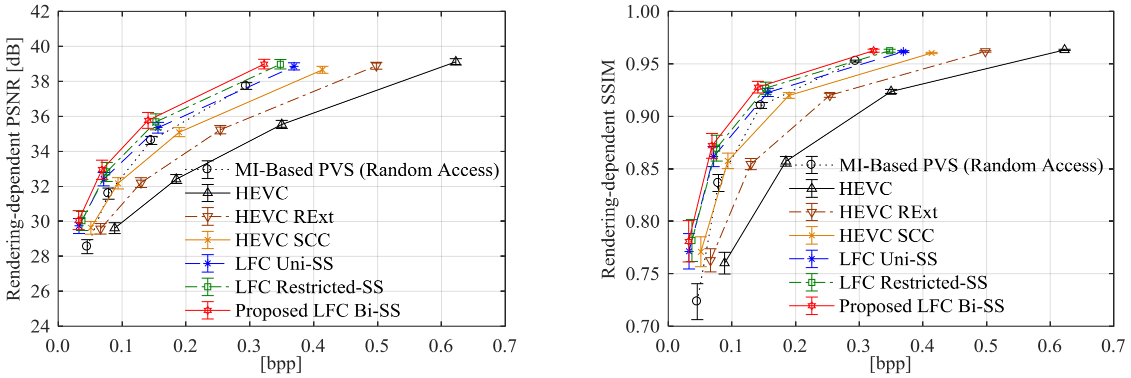
(with respect to HEVC) and 0.56 dB and 34.17 % (with respect to the LFC Uni-SS solution).

- **Regarding Different Objective Quality Metrics** – Comparing the different objective quality metrics (in Tables 4.5 and 4.6, and Figures 4.7 to 4.14), it can be seen that there is a consistent relative RD performance between the tested solutions along all metrics. Moreover, in all cases, the proposed LFC Bi-SS outperforms all of the other tested benchmarks.
- **Regarding Different Acquisition Parameters and Scene Characteristics** – Comparing the results presented in Figures 4.7 to 4.14, it is possible to observe that the performance of the proposed LFC Bi-SS seems to be more related to the used acquisition setup than to the scene type (see information about the different LF acquisition setup in Appendix A for each LF test image).



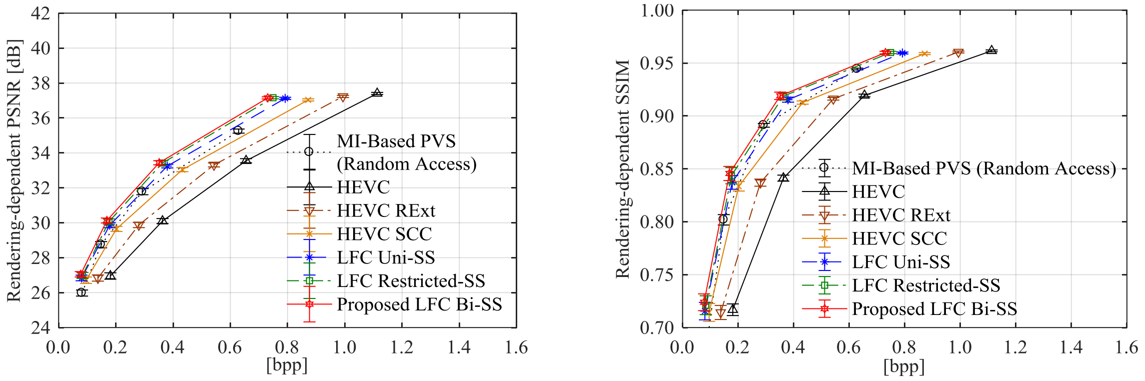
(a) *(b)*

Figure 4.7 RD performance for Fredo image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



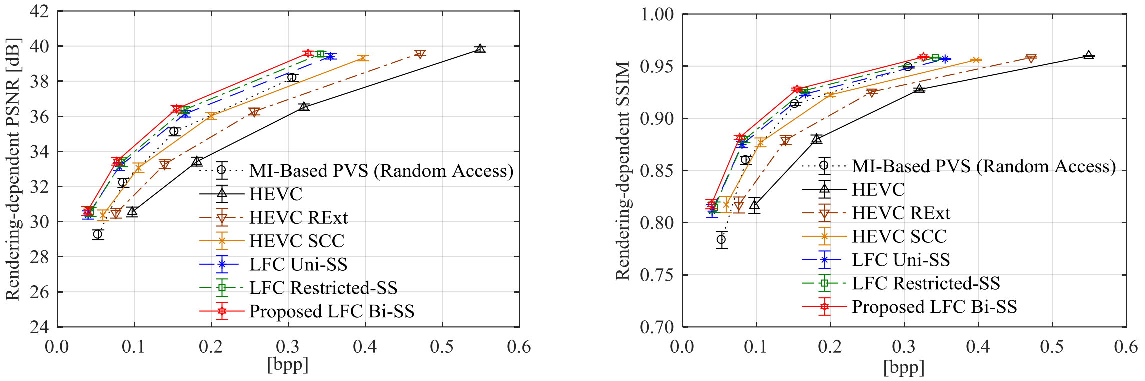
(a) *(b)*

Figure 4.8 RD performance for Seagull image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



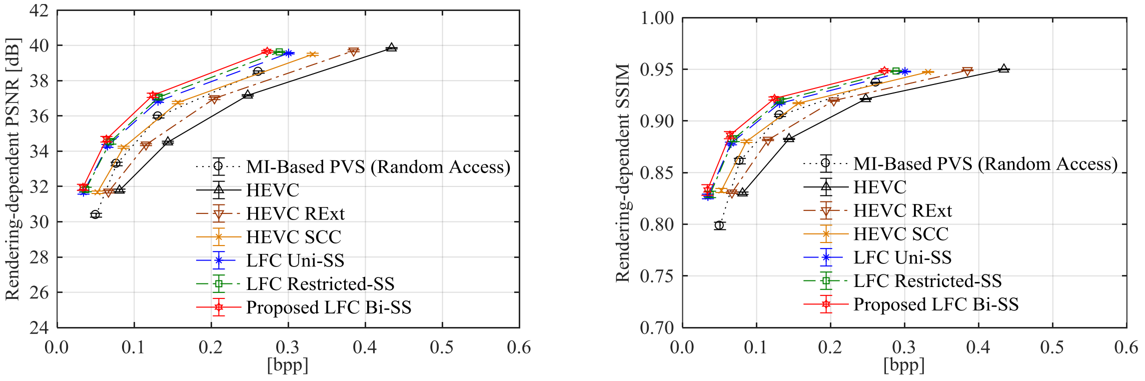
(a) *(b)*

Figure 4.9 RD performance for Laura image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



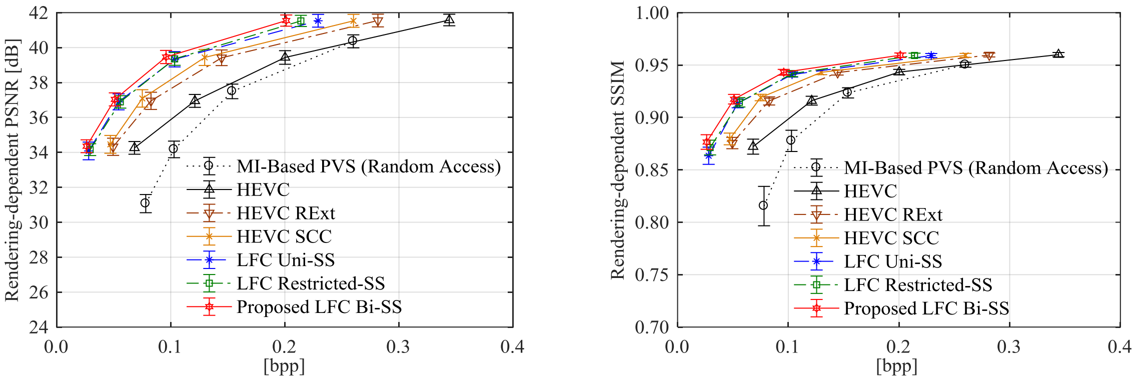
(a) *(b)*

Figure 4.10 RD performance for Jeff image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



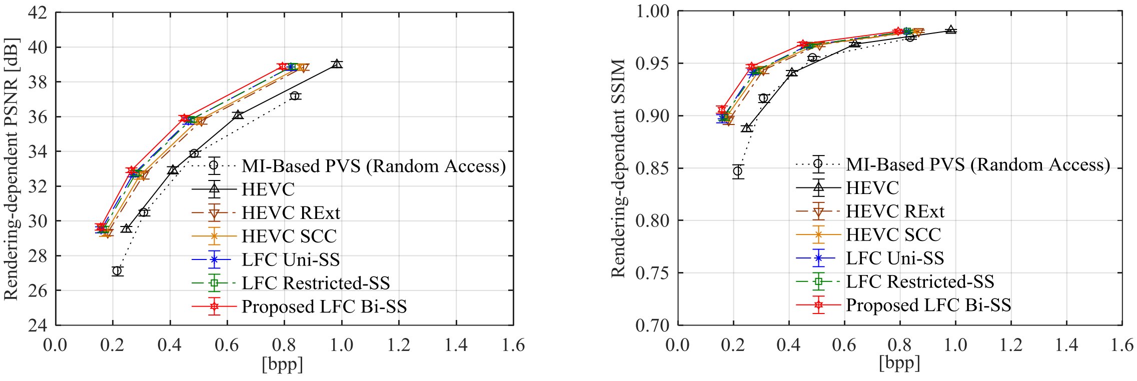
(a) *(b)*

Figure 4.11 RD performance for Zhengyun1 image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



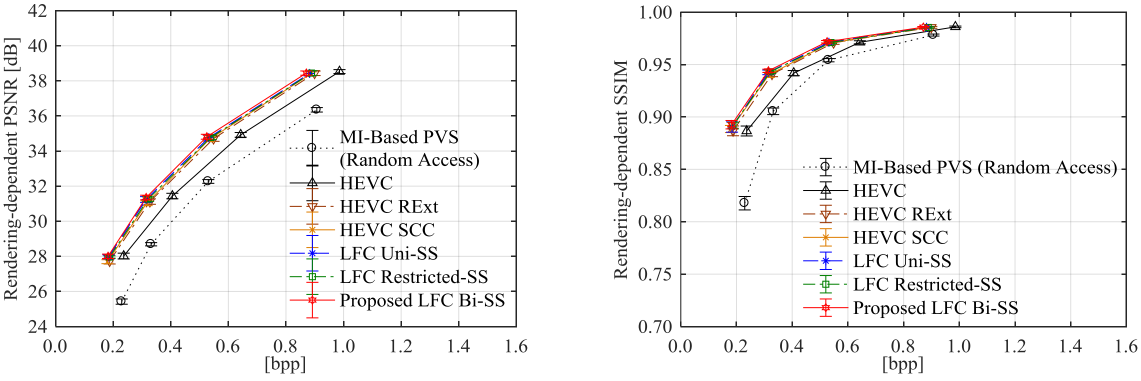
(a) *(b)*

Figure 4.12 RD performance for Demichelis Spark (frame1) image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



(a) *(b)*

Figure 4.13 RD performance for Plane and Toy (frame 123) image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values



(a) *(b)*

Figure 4.14 RD performance for Robot 3D (frame 54) image considering: (a) Rendering-dependent PSNR versus bpp values; and (b) Rendering-dependent SSIM versus bpp values

4.5.3 Visual Quality Evaluation

For a visual inspection of views rendered from coded LF images, Figures 4.15 and 4.16 illustrate the central view rendered from two LF images *Demichelis Spark (frame 1)* (left), and *Zhengyun1* (right) for two different bpp values.

The quality of the original LF image (see Figures 4.15a and 4.16a) is then compared to the LF image coded with the proposed LFC Bi-SS (see Figures 4.15b and 4.16b), LFC Restricted-SS (see Figures 4.15c and 4.16c), and HEVC (see Figures 4.15d and 4.16d).

From these results, it is possible to conclude that the proposed LFC Bi-SS presents considerably better visual quality than HEVC and improvements are also noticeable when compared to the LFC Restricted-SS solution. For instance, in the hand and face of *Demichelis Spark (frame 1)* and in the eyes and cheeks of *Zhengyun1*.

4.5.4 Influence of MI Cross Correlation

In order to analyze the influence of the MI cross-correlation in the performance of the proposed LFC Bi-SS solution, two different LF images are used to illustrate the case where the MI cross-correlation is differently distributed in a neighborhood. Notably, the second frame of *Plane and Toy* sequence (i.e., frame 23 depicted in Figure A.8a) is used to exemplify the case where the MI cross-correlation varies for the same camera parameters due to the different distance of the main object relatively to the camera [54]. In this case, *Plane and Toy (frame 23)* presents a more rapid decrease in the MI cross-correlation in a neighborhood compared to *Plane and Toy (frame 123)*, as illustrated in Figure A.8. This analysis aims also at studying how close to the experimental results are the theoretical conclusions drawn in [51, 211] and the approximations considered in Section 4.4.1.

For this, Table 4.7 shows the BD performance of the proposed LFC Bi-SS against the LFC Uni-SS (where only uni-prediction is used) for the two abovementioned LF images. Moreover Table 4.8 summarizes some statistics of relevant encoding results for ‘high bpp’ values, such as: percentages of prediction modes usage, and SS bi-prediction usage.

Analyzing the statistics for the LFC Bi-SS solution (see Table 4.8), and comparing the results for *Plane and Toy (frame 123)* and *Plane and Toy (frame 23)*, it can be observed that the percentage of usage of bi-prediction is larger than uni-prediction in the case where the MI cross-correlation decreases more rapidly (*frame 23*). In addition to this, comparing the statistics in Table 4.8 with the BD results for ‘high bpp’ values in Table 4.7, it can be seen that the usage of bi-prediction in the LFC Bi-SS solution is the main reason for the achieved coding gains against LFC Uni-SS solution. Moreover, the larger the percentage of usage of bi-prediction (in Table 4.8), the greater the BD gains against the LFC Uni-SS solution will be (in Table 4.7).

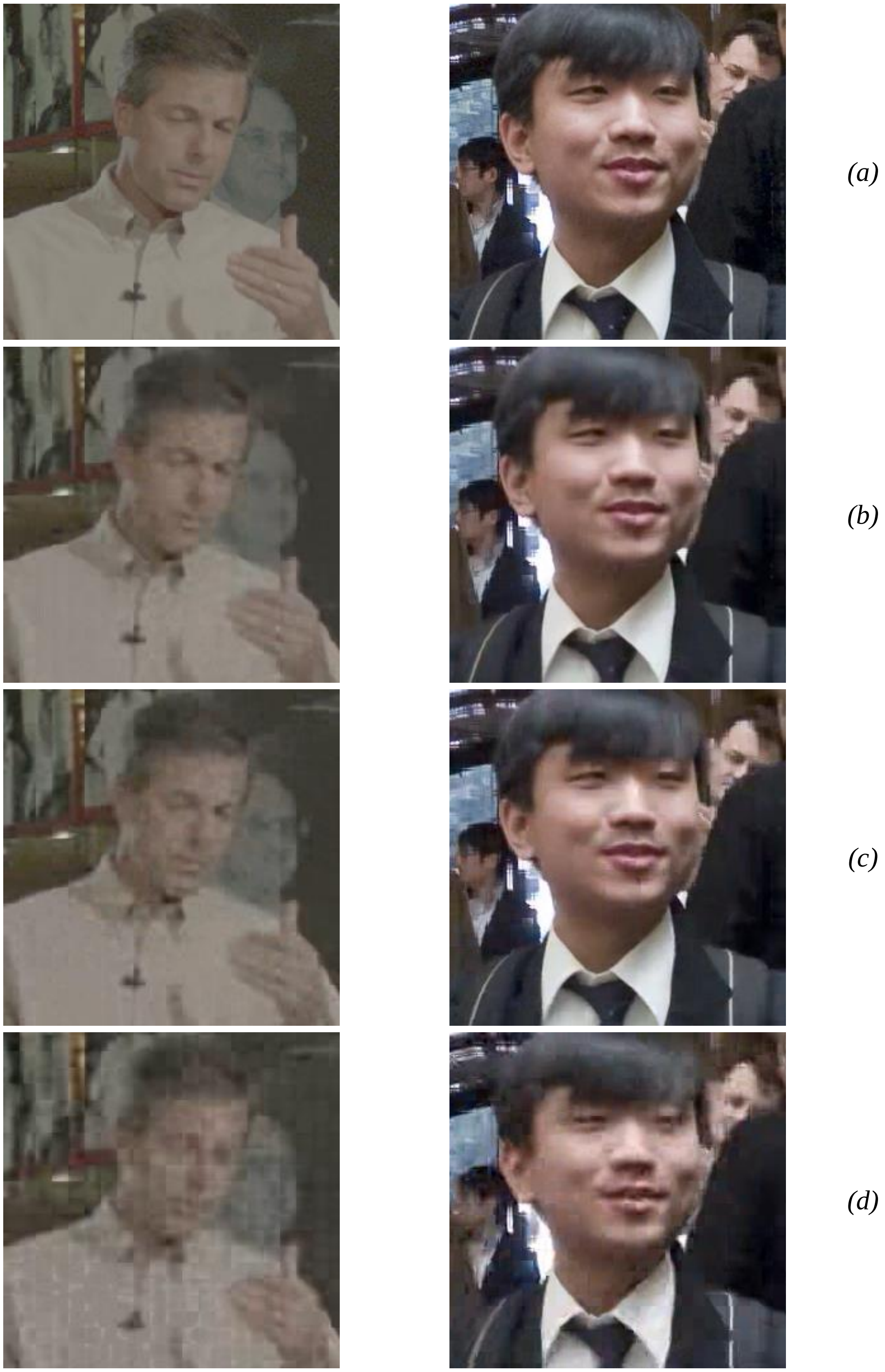


Figure 4.15 Comparison of a portion from the central view rendered from: (a) original image; (b) compressed image using the proposed LFC Bi-SS solution; (c) compressed image using LFC Restricted-SS; and (d) compressed image using HEVC. The compressed results are shown for the LF images (left to right): Demichelis Spark at 0.05 bpp; and Zhengyun1 at 0.06 bpp

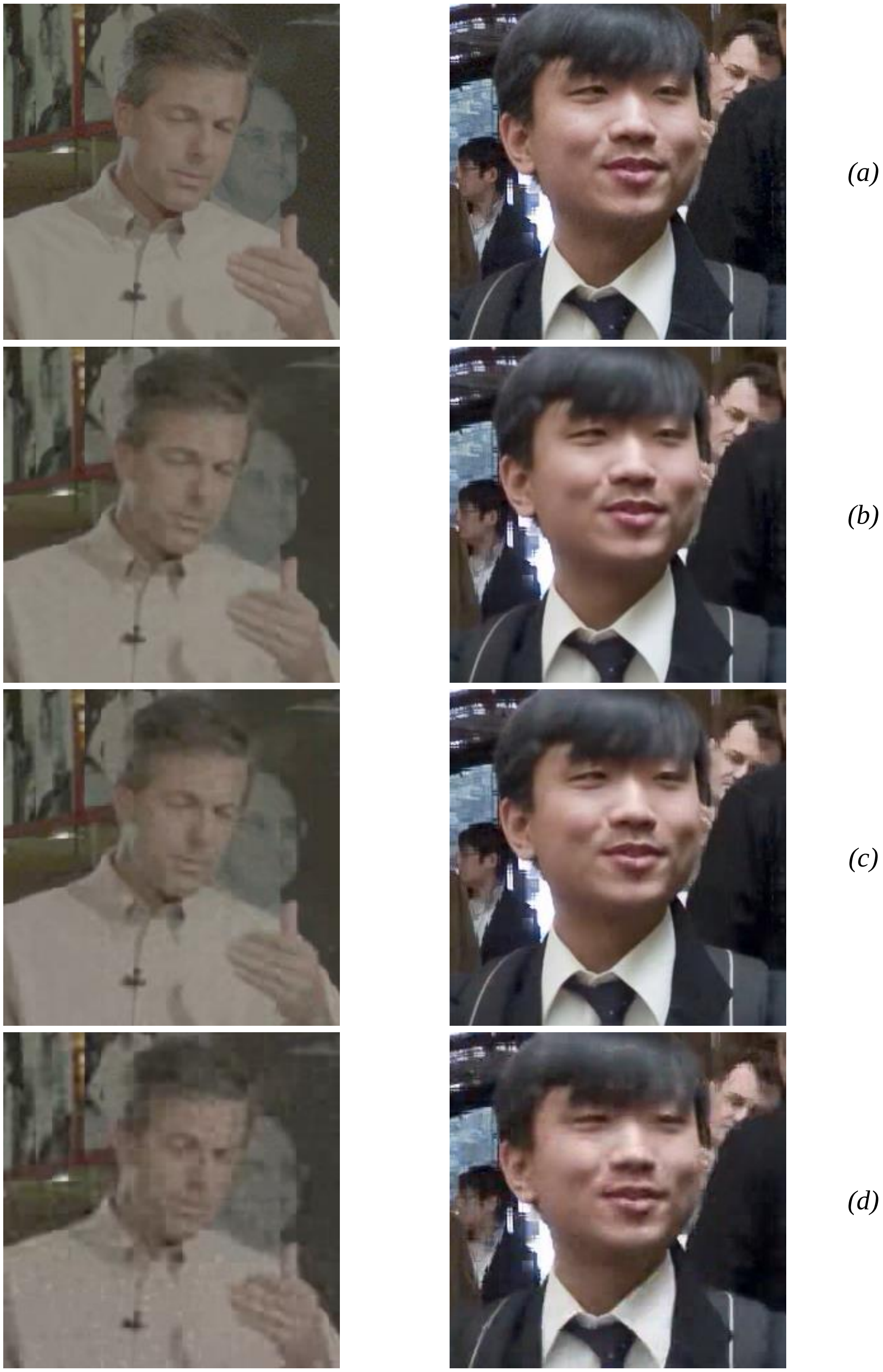


Figure 4.16 Comparison of a portion from the central view rendered from: (a) original image; (b) compressed image using the proposed LFC Bi-SS solution; (c) compressed image using LFC Restricted-SS; and (d) compressed image using HEVC. The compressed results are shown for the LF images (left to right): Demichelis Spark at 0.09 bpp; and Zhengyun1 at 0.12 bpp

Table 4.7 Influence of MI cross-correlation in RD performance (using BD metrics [210]*) of the for 'high bpp' and 'low bpp'*

Performance for 'high bpp' values (QP 22, 27, 32, 37)		
Plane and Toy	**LFC Uni-SS vs. LFC Bi-SS**	
	PSNR [dB]	**BR [%]**
Frame 123	0.28	-4.51
Frame 23	0.36	-6.13
Performance for 'low bpp' values (QP 37, 32, 37, 42)		
Plane and Toy	**LFC Uni-SS vs. LFC Bi-SS**	
	PSNR [dB]	**BR [%]**
Frame 123	0.30	-4.89
Frame 23	0.41	-6.50

Table 4.8 Influence of MI cross-correlation in mode selection statistics for 'high bpp' (QP 22)

Plane and Toy	**Statistics for LFC Uni-SS**			**Statistics for LFC Bi-SS**				
	Prediction Modes			**Prediction Modes**			**SS Prediction**	
	Intra	**SS**	**SS-skip**	**Intra**	**SS**	**SS-skip**	**Uni**	**Bi**
Frame 123	59.24%	40.24%	0.52%	57.18 %	42.29 %	0.53 %	60.36 %	39.64 %
Frame 23	69.66%	29.42%	0.93%	67.77 %	30.90 %	1.33 %	54.24 %	45.76 %

Table 4.9 Prediction mode selection statistics for 'low bpp' values (QP 42)

Plane and Toy	**Statistics for LFC Uni-SS**			**Statistics for LFC Bi-SS**				
	Prediction Modes			**Prediction Modes**			**SS Prediction**	
	Intra	**SS**	**SS-skip**	**Intra**	**SS**	**SS-skip**	**Uni**	**Bi**
Frame 123	31.08%	46.17%	22.74%	30.49 %	46.22 %	23.29 %	87.14 %	12.86 %
Frame 23	44.53%	31.88%	23.59%	43.72 %	30.51 %	25.77 %	74.95 %	25.05 %

Finally, it is possible to see that the theoretical hypotheses from Section 4.4.1 are consistent with the obtained experimental results.

4.5.5 Impact of Quantization Noise

Table 4.9 also illustrates the statistical results also for 'low bpp' values (corresponding to QP value 42) so as to analyze the performance of the Bi SS prediction when the SS reference degrades.

Comparing the results in Tables 4.8 and 4.9 for all test images, it can be observed that using higher QP values results in increasing percentages of usage of the SS uni-prediction, as well as SS-skip modes. This is due to the fact that the Lagrangian multiplier, in (4.1) and Figure 4.6, increases with increasing QP values [52] and, for this reason, the possible quality

Table 4.10 LFC BI-SS encoding complexity regarding LFC Uni-SS and LFC Restricted-SS with QP value set to 32

Image	**LFC Bi-SS Encoding Time [s]**	**LFC Bi-SS Encoding Time Ratio:**	
		LFC Uni-SS	**LFC Restricted-SS**
Fredo	27836	2.01	1.66
Seagull	30665	1.93	1.7
Laura	34479	2.07	1.89
Jeff	28360	2.01	1.69
Zhengyun1	30604	1.93	1.70
Demichelis Spark (frame 1)	2552	2.20	1.68
Plane and Toy (frame 123)	1430	2.26	1.84
Robot3D (frame 54)	1670	2.25	1.77
	Average	**2.08**	**1.74**

improvements of using the bi-prediction do not justify the higher amount bits needed for transmitting the SS vectors when minimizing the Lagrangian cost.

4.5.6 Computational Complexity

The significantly better performance of the LFC Bi-SS solution comes with the price of additional computational load compared to LFC Uni-SS and LFC Restricted-SS.

To illustrate this fact, Table 4.10 presents the encoding time for each LF image of the LFC Bi-SS solution and also the encoding time ratio, in terms of $EncTime_{\text{Bi-SS}}/EncTime_{\text{bench.}}$, with respect to the LFC Uni-SS and LFC Restricted-SS benchmark solutions. For this, encoding times were obtained using a machine with Intel Xeon E5-2620 v2 clocked at 2.10 GHz and using gcc 4.8.3 compiler.

As can be seen, although the LFC Bi-SS is 2.08 times slower than LFC Uni-SS and 1.74 times slower than LFC Restricted-SS (in average), the additional gains of using the LFC Bi-SS (up to 0.65 dB and 17.32 % of bit savings compared to LFC Uni-SS, and 0.50 dB and 9.32 % of bit savings compared to LFC Restricted-SS) are relevant.

Nevertheless, note that a fast search approach can still be adopted, for instance, by taking advantage of the regular SS vector distribution depicted in Figure 4.4. However, since this kind of fast coding solutions does not typically have normative restrictions, it was not considered in this Thesis.

4.6 Final Remarks

This chapter proposed an efficient LFC solution based on HEVC and using the SS compensated prediction concept. Further improvements were also proposed to achieve better RD performance to the proposed LFC solution. Notably, a novel SS vector prediction was proposed, named MIVP, which explored the particular characteristics of the SS prediction data. Moreover, an improved Bi-SS prediction was also proposed, in which two predictor blocks were jointly estimated from the same search window by using a locally optimal rate-constrained algorithm, instead of independently deriving them in different areas of the LF image, as previously proposed in the literature.

As discussed in this chapter, the proposed HEVC-based LFC coding architecture was shown to be advantageous in terms of the simplicity of the coding format, which is less dependent on a very precise LF camera calibration process, while keeping the encoder/decoder complexity and memory load comparable to HEVC. In addition, the proposed LFC Bi-SS (using MIVP) led to significantly superior performance when compared to HEVC Still Picture Profile, presenting gains of up to 4.29 dB and 61.09 % of bit savings. Furthermore, jointly estimating the two candidate blocks for Bi-SS prediction led to further RD improvements when compared to the case where only one candidate block is estimated (with RD gains of up to 0.65 dB and 17.32 % with respect to the LFC Uni-SS solution), as well as compared to the case where to the two candidate blocks are independently estimated (with RD gains of up to 0.50 dB and 9.32 % with respect to the LFC Restricted-SS solution). Moreover, it was seen that significantly better RD performance compared to HEVC Still Picture Profile was also observed for LF images captured with the Lytro Illum traditional LF camera.

Chapter 5

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. Dealing with this specific concern is an essential requirement for enabling faster deployment of new LF imaging application services in the consumer market, and, for this reason, the main focus of this chapter will be to propose LF coding solutions aiming at providing backward display compatibility.

An LF data format that is backwardly compatible with existing receivers (either 2D, or current stereo or multiview) would also allow a more gradual introduction of LF content for end-users, who can have different preferences and can be using different display devices. As illustrated in Figure 5.1, this would mean that a legacy 2D (or a legacy stereo or multiview) consuming device that does not explicitly support LF content would be able to display a 2D (or stereo or multiview) version of the captured LF content, while a more advanced future LF consuming device would be able to display the full LF content.

For enabling this, an efficient scalable LF coding approach is then desirable (as seen Figure 5.1), where by decoding only the adequate subsets of the scalable bitstream, 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, the contribution of this chapter is twofold:

1) First, a display scalable architecture for LF content coding is proposed 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 (see Figure 5.1); to the end-user who wants have a more immersive and interactive visualization by using a more advanced LF display technology (see Figure 5.1), such as

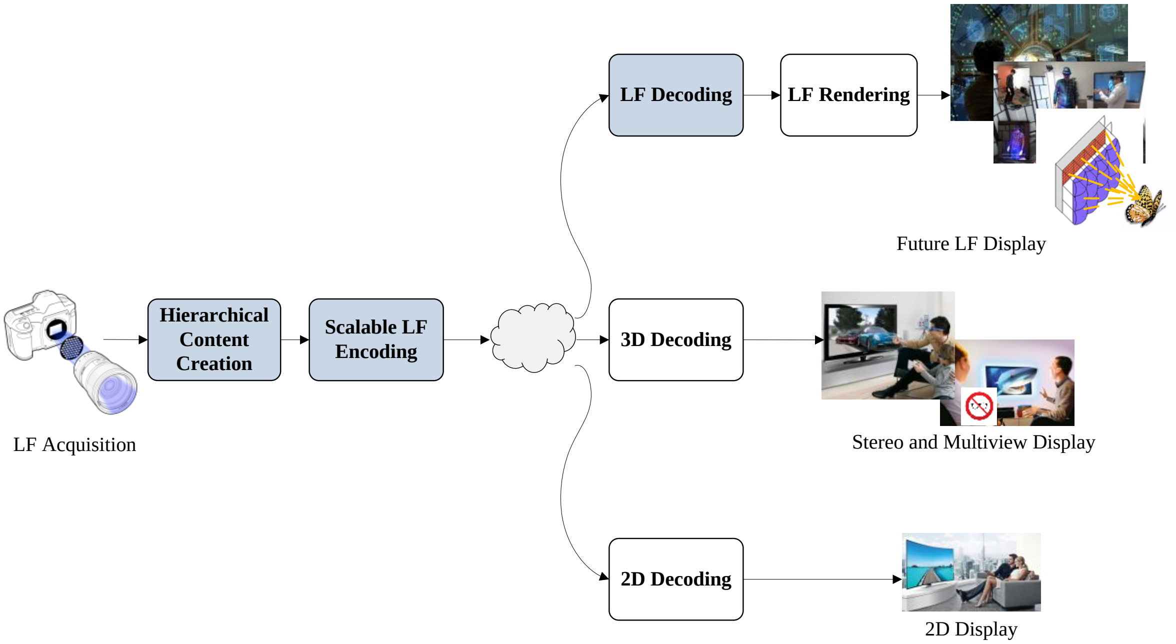


Figure 5.1 Target application scenario for the proposed scalable LF coding solution with backward display compatibility

an integral imaging display [5, 14–16] or a head mounted display for augmented and virtual reality [23, 24].

2) Based on this hierarchical coding architecture, an LF enhancement codec is also proposed to efficiently encode the LF content in the highest layer. For this, the SS compensated prediction, which has been proposed in Chapter 4, is here combined with a novel Inter-Layer (IL) prediction scheme for improving the RD coding performance compared to independent compression of the three different layers (i.e., the simulcast case). The proposed IL prediction mechanism aims at exploiting the existing redundancy between the multiview and the LF content. To accomplish this, a prediction picture is built and used as a new reference frame in an IL compensated prediction scheme.

The remainder of this chapter is organized as follows: Section 5.1 proposes a scalable LF coding solution architecture for backward display compatibility (as illustrated in Figure 5.1); Section 5.2 briefly overviews the process for generating the content for each hierarchical layer (see Figure 5.1); Section 5.3 presents the proposed LF enhancement layer coding solution; while Section 5.4 presents the proposed novel IL prediction construction to be used for further improving the RD performance; Section 5.5 performs the evaluation of the proposed scalable codec; and, finally, Section 5.6 concludes the chapter.

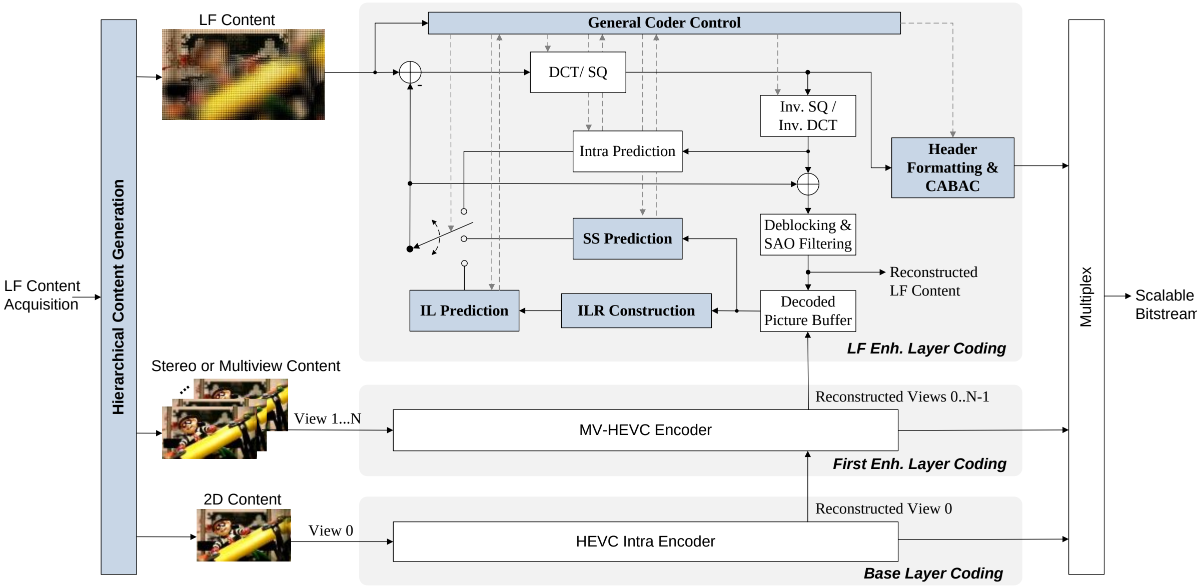


Figure 5.2 Scalable LF coding architecture using three hierarchical layers for backward display compatibility. The novel and modified blocks are highlighted in blue shaded blocks.

5.1 Proposal for a Scalable Coding Architecture with Backward Display Compatibility

A display scalable architecture for LF coding with a three-layer approach is proposed here and illustrated in Figure 5.2. As can be seen, each layer of this scalable coding architecture represents a different level of display scalability:

- **Base 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.
- **Stereo or Multiview Enhancement Layer (First Enhancement 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.
- **LF Enhancement Layer (Second Enhancement Layer)** – This layer represents the additional data needed to support full LF content display.

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

Therefore, the coding information flow in the proposed Display Scalable Light Field Coding (DS-LFC) solution is defined as the following:

1) In the base layer, a 2D view is coded with conventional HEVC [2] 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 Figure 5.2.
2) The content in the first enhancement layer can be encoded by using a standard stereo or multiview coding solution [39, 40, 132, 213], and the reconstructed 2D views are then made available to be used for coding of the LF enhancement layer. For the work presented in this chapter, the multiview extension of HEVC, MV-HEVC [40], is adopted. With these solutions [39, 40, 132, 213], inter-view prediction can be used to improve the coding efficiency between the base layer and the first enhancement 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 [39], and in the 3D video coding extensions of the HEVC [40]. For a good review of these 3D video coding solutions, the reader can refer to [39, 40, 132, 213].
3) Finally, the content in the LF enhancement layer, i.e., the LF content, is encoded by using the LF enhancement coding solution proposed in Section 5.3.

5.2 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 microlens array structure (i.e., the array packing scheme and the microlens shape) – needs to be known when encoding and decoding the highest scalable layer.

In the work presented in this chapter, the two rendering algorithms proposed in [10] and referred to as Basic Rendering and Weighted Blending are adopted for this hierarchical content generation. As previously presented in Chapter 2, the idea behind these algorithms is to combine suitable patches from each MI to properly compose a 2D view image. Then, as explained in [10], the process of generating a 2D view image can be controlled by the following two main parameters (see Chapter 2):

- **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.

5.3 Proposal for an Efficient LF Enhancement Layer Coding Solution

Since the lower layers of the proposed DS-LFC codec presented in Section 5.1 are based on the HEVC [2] 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 Figure 5.2, so as to modify as few aspects of the underlying architecture as possible. Notice that, although the LF enhancement layer encoder presented in Figure 5.2 targets LF still image coding, it can be also extended for scalable LF coding by including also the HEVC inter-frame coding.

The main blocks of the proposed HEVC-based LF enhancement encoder (as depicted Figure 5.2) are explained in the following.

5.3.1 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 (Figure 5.2). The decision between the different available prediction modes is made in an RDO manner, [52] as in conventional HEVC [2].

5.3.2 SS Compensated Prediction

Since the LF content in the highest enhancement layer presents a significant cross-correlation between neighbor MIs in the 2D grid (as discussed in Chapter 2), the SS compensated prediction (see Figure 5.2), which has been proposed in Chapter 4, can be also used as an alternative to exploit the existing redundancy within the LF image and to improve coding efficiency. For this to be possible, the SS reference is made available in the reference picture list(s) for intra-coded frames. As a result of the SS prediction, the residual information and the SS vector(s) are coded and sent to the decoder.

5.3.3 IL Compensated Prediction

In addition, 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 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 Figure 5.2) when encoding the LF image. The process for constructing the ILR picture is proposed in Section 5.4.

5.3.4 Extending HEVC Syntax for Scalable LF Content Coding

Further extending HEVC to support scalable LF coding basically means to introduce the IL and SS references into the reference picture list(s) of HEVC and to allow them to be used by the existing HEVC inter prediction modes. Consequently, no changes in the lower levels of the syntax and decoding processes of HEVC are needed, and it only involves including some additional high level syntax elements to the HEVC.

Notably, the following two important pieces of information are carried through the HEVC high level parameter sets in order to be available at the decoder side:

1) **Acquisition Parameters** – As discussed in Section 5.2, this information is composed of a set of 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, microlens structure, size and position of the patches).
2) **LF Dependency Information** – Since two new references are included into the HEVC reference picture list(s), this LF dependency information is necessary to indicate which LF references (among IL and SS references) are available for each of the two reference picture lists (similarly to HEVC). This way, it is possible to distinguish these new reference frames from each other (as well as from temporal references).

The LF dependency information is used to build the reference picture lists for a LF image being coded. In this case, both the ILR and SS reference are appended to the HEVC reference picture lists, becoming available for prediction of the current picture.

It is also important to notice that, in terms of lower syntax levels (e.g., CU level of the HEVC [2]), the decoding modules do not need to be aware of the reference picture type, and the distinction. Consequently, the encoder can select the best reference picture in an RDO sense, resulting only in an index of the position in the reference picture list. This index is conveyed along with the prediction information and transmitted to identify, at the decoder side, the position of the used reference picture.

5.4 Proposal for a Novel Inter-Layer Prediction Construction

As discussed in the previous section, higher coding efficiency is likely to be achieved by exploring the existing redundancy between the various layers by means of the proposed IL prediction scheme. This IL prediction scheme builds an ILR picture which is then used to predict the LF image being coded. 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 Figure 5.2;
- **Acquisition Parameters** – These parameters comprise information from the LF capturing process (such as the MI resolution and the microlens structure) and also information from the 2D view generation process (i.e., size and position of the patches). As explained in Section 5.3.4, this information has to be conveyed along with the bitstream to be available at the decoding side.

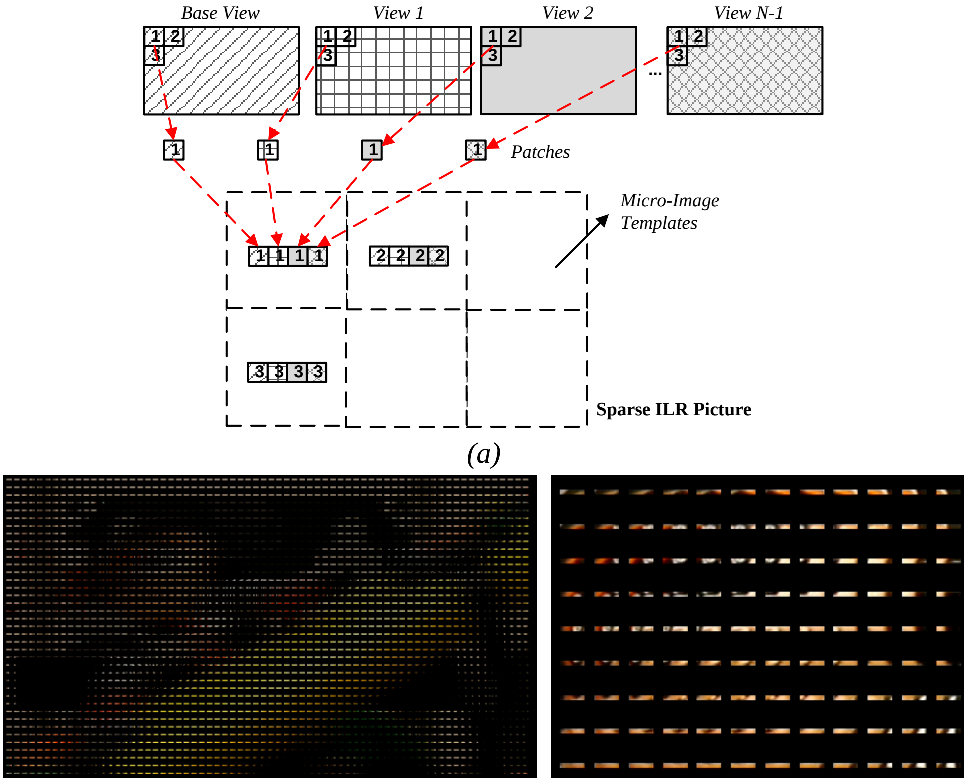
Therefore, two steps are distinguished when generating an ILR picture, which are explained in the following:

1) **Patch Remapping** – Although most of the LF information is discarded when rendering each view in the hierarchical layer generation block in Figure 5.2, 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.
2) **MI Refilling** – Afterwards, this step aims at emulating the significant cross-correlation existing between neighboring MIs so as to synthesize the LF information that was discarded in hierarchical layer generation block.

5.4.1 Patch Remapping

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 Basic Rendering algorithm (see Chapter 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 Figure 5.3a. A template for the LF image assembles all patches, and the output is referred to as the sparse ILR picture, as seen in Figure 5.3b.



(a)

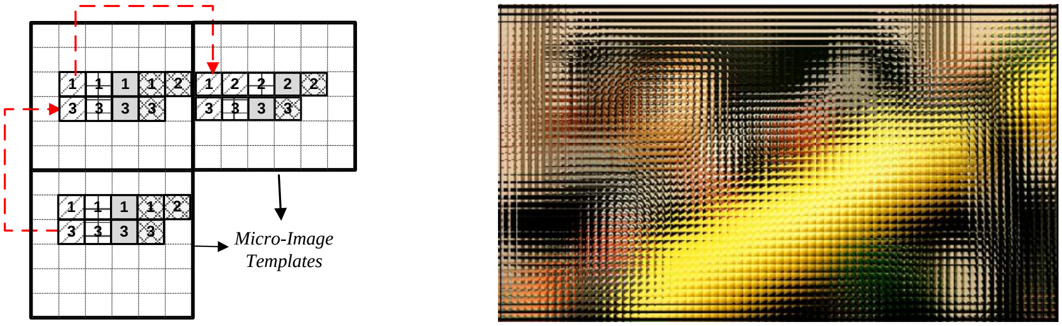
(b)

Figure 5.3 The Patch Remapping step to generate a sparse ILR picture for the Plane and Toy (frame 123) shown in Figure A.8b: (a) Algorithm walkthrough; (b) Example of a sparse ILR picture (left) and its enlargement showing the area of the toy's face (right)

5.4.2 MI Refilling

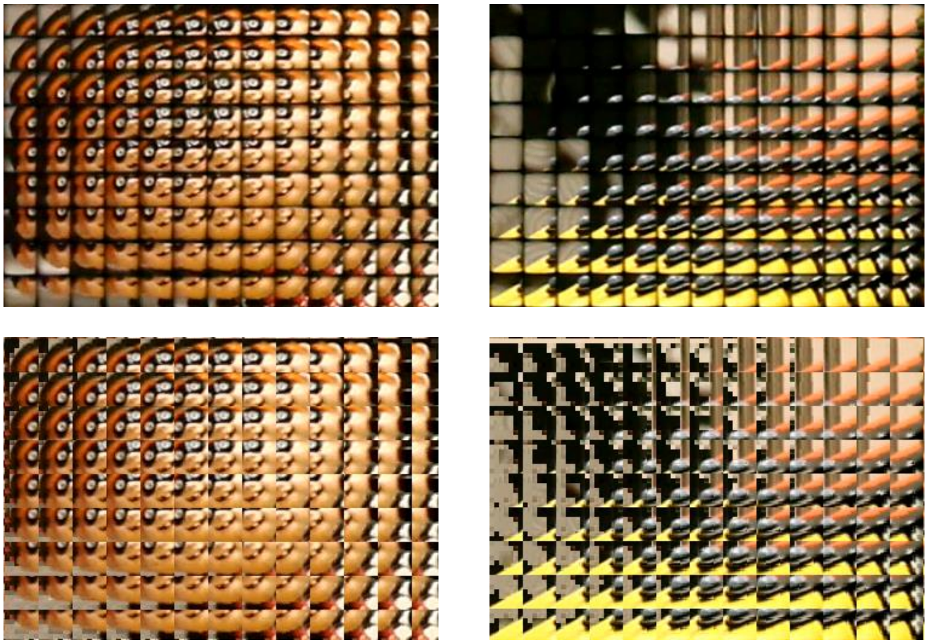
The input for this step is the sparse ILR picture generated by the Patch Remapping. Basically, the MI Refilling aims at emulating the significant cross-correlation existing between neighboring MIs so as to fill the holes in the sparse ILR picture (see Figure 5.3b) as much as possible.

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 5.2). An illustrative example of this process is shown in Figure 5.4a 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 Figure 5.4a) 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 Figure 5.4b).



(a) *(b)*

Figure 5.4 The MI Refilling step to generate an ILR picture for the Plane and Toy (frame 123) shown in Figure A.8b: (a) Algorithm walkthrough; and (b) Example of an ILR picture



(a) *(b)*

Figure 5.5 Comparison of a portion from the original LF image Plane and Toy (frame 123) (top) and the corresponding constructed ILR (bottom) for a particular patch size. The ILR construction is more accurate for areas of the rendered image where the selected patch size is focused on (a) than the remaining out of focus areas (b) of the rendered image

It is worthwhile to notice that a characteristic of the MI refilling process is that the constructed ILR picture is considerably more accurate in areas where the used rendered 2D views are in focus compared to out of focus areas. This is illustrated in Figure 5.5 for in focus (see Figure 5.5a) and out of focus (see Figure 5.5b) areas when using the patch size 4 for the LF image *Plane and Toy (frame 123)* shown in Figure A.8b. As can be seen, since the actual disparity is not known in the out of focus areas and, consequently, is assumed to be given by the patch size, the reconstruction is less accurate than in the in focus areas. Therefore, the quality of the constructed ILR picture is likely to improve as more areas of the rendered views are in focus.

Moreover, 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.

5.5 Performance Assessment

This section assesses the performance of the proposed DS-LFC codec. For this purpose, the test conditions and tested coding solutions are firstly introduced and, then, the obtained results are presented and discussed.

5.5.1 Test Conditions

The test conditions that are adopted to assess the performance of the proposed DS-LFC solution can be summarized as follows.

5.5.1.1 Test Images and Hierarchical Content Generation Parameters

Six LF images with different spatial and MI resolutions are considered so as to achieve representative RD results. These are (see Appendix A): (i) *Fredo*, as shown in Figure A.1; (ii) *Seagull* (Figure A.4); (iii) *Laura* (Figure A.3); (iv) *Demichelis Spark* (first frame of the sequence shown in Figure A.7); (v) *Robot 3D* (frame number 54 of the sequence shown in Figure A.9); and (vi) *Plane and Toy* (frame number 123 of the sequence shown in Figure A.8b).

The complete characterization of these LF images is shown in Tables A.1-A.5. The original LF images were rectified so as to have all MIs with integer number of pixels, and they were then converted to the Y'CbCr 4:2:0 color format.

To generate the content for the 2D, stereo or multiview layers, the six LF test images were processed using two different algorithms: i) Basic Rendering [10], and ii) Weighted Blending [10] algorithms (see Chapter 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 5.1.

5.5.1.2 Coding and Evaluation Conditions

For generating the experimental results, the following coding conditions were tested:

Table 5.1 Test conditions adopted for the proposed DS-LFC solution, notably, the patch sizes and positions for generating content for the lower hierarchical layers

LF Image	Patch Size (Focus Plane)	Patch Positions (View's Perspectives)
Fredo	10	{(-24,0), (-18,0), (-12,0), (-6,0), (0,0), (6,0), (12,0), (18,0), (24,0)}
Seagull	9	{(-24,0), (-18,0), (-12,0), (-6,0), (0,0), (6,0), (12,0), (18,0), (24,0)}
Laura	10	{(-24,0), (-18,0), (-12,0), (-6,0), (0,0), (6,0), (12,0), (18,0), (24,0)}
Demichelis Spark	12	{(-8,0), (-6,0), (-4,0), (-2,0), (0,0), (2,0), (4,0), (6,0), (8,0)}
Robot 3D	4	{(-8,0), (-6,0), (-4,0), (-2,0), (0,0), (2,0), (4,0), (6,0), (8,0)}
Plane and Toy	4	{(-8,0), (-6,0), (-4,0), (-2,0), (0,0), (2,0), (4,0), (6,0), (8,0)}

- **Codec Software Implementation** – For these tests, the reference software for the MV-HEVC extension version 12.0 [214] is used as the base software for implementing the proposed DS-LFC codec.

- **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.

- **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.

- **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 [210] 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.

- **Alternative 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 Rendering-dependent PSNR (as in (4.8)) and Rendering-dependent SSIM (as in (4.9)) metrics. To have a representative number of rendered views, a set of 9 views are rendered from viewpoint positions equally distributed in horizontal and vertical directions. 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 [10]).

- **Visual Quality Evaluation** – Moreover, a visual inspection of the quality of central views rendered from the reconstructed LF image is also considered for assessing the proposed DS-LFC solution. For rendering the views, the same algorithm used for generating the content in the lower layers is adopted (i.e., Basic Rendering or Weighted Blending [10]). For all compared coding solutions, the QP values of the encoder is adjusted to lead to the same bpp value.

5.5.1.3 Tested Solutions

The performance of the proposed DS-LFC solution is compared to the following solutions:

- **VI-Based PVS (Low Delay P)** – This solution represents a benchmark coding approach for providing display scalability. For instance, each VI could be seen as a different view and the LF content is represented in a multiview format. Then, as in MV-HEVC, each view is encoded in a different layer to support backward compatibility and scalability. In this specific test, since the MV-HEVC reference software [214] does not support more than 64 views, HEVC is rather adopted where inter-view prediction is allowed using a PVS-based approach (as seen in Chapter 3). Therefore, similarly to what has been proposed in [92, 175], a PVS of VIs is coded using HEVC with the Low Delay P [176] configuration. However, in order to fairly compare this solution with the coding of views in the lower layers of the proposed DS-LFC solution, the QP values are kept the same for all VIs in the PVS. Various VI scanning orders were tested (i.e., raster, parallel, zigzag, and spiral), but only the spiral order is presented as it achieved the best RD performance.

- **VI-Based PVS (Random Access)** – In this case, the PVS of VIs scanned in spiral order is encoded using HEVC with the Random Access [176] configuration. Similarly to the previous solution (VI-based PVS (Low Delay P)), the largest CB size is set to 16×16 and the QP values are kept the same for all VIs in the PVS.

- **HEVC (Single Layer)** – In this case, the entire LF image is encoded into a single layer with HEVC using the Main Still Picture profile [2]. 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.

- **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 [176].

- **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

Table 5.2 RD performance (using BD metrics [210]*) of the proposed DS-LFC codec with respect to other coding architectures (QP values 22, 27, 32, and 37)*

LF Image	**VI-based PVS (Low Delay P)**		**VI-based PVS (Random Access)**		**HEVC (Single Layer)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
Fredo	8.67	-80.99	8.46	-78.74	2.08	-32.27
Seagull	5.64	-75.32	5.33	-73.54	**2.40**	**-37.90**
Laura	6.15	-66.95	5.42	-62.71	1.32	-19.99
Demichelis Spark (frame 1)	4.43	-66.37	3.08	-52.55	-0.19	6.80
Robot 3D (frame 54)	6.83	-58.75	5.23	-51.74	**-0.56**	**7.54**
Plane and Toy(frame 123)	5.82	-60.27	3.84	-46.92	0.32	-5.13
Average	**6.26**	**-68.11**	**5.23**	**-61.03**	**0.90**	**-13.49**

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.

In the case of the proposed DS-LFC codec depicted in Figure 5.2, 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.

5.5.2 Overall DS-LFC RD Performance Evaluation

To assess the performance of the proposed DS-LFC codec, Tables 5.2 and 5.3 present the BD-PSNR and BD-BR [210] results with respect to the benchmark solutions for the six LF test images presented in Section 5.5.1.1. Additionally, Figure 5.6 shows the RD performance in terms of the luma PSNR of the image in the LF enhancement layer versus the total bitrate for encoding all hierarchical layers (in bpp). For the results in Tables 5.2 and 5.3, as well as in Figure 5.6, the Basic Rendering algorithm was adopted for generating the content in the lower hierarchical layers.

From these results, the following conclusions can be derived:

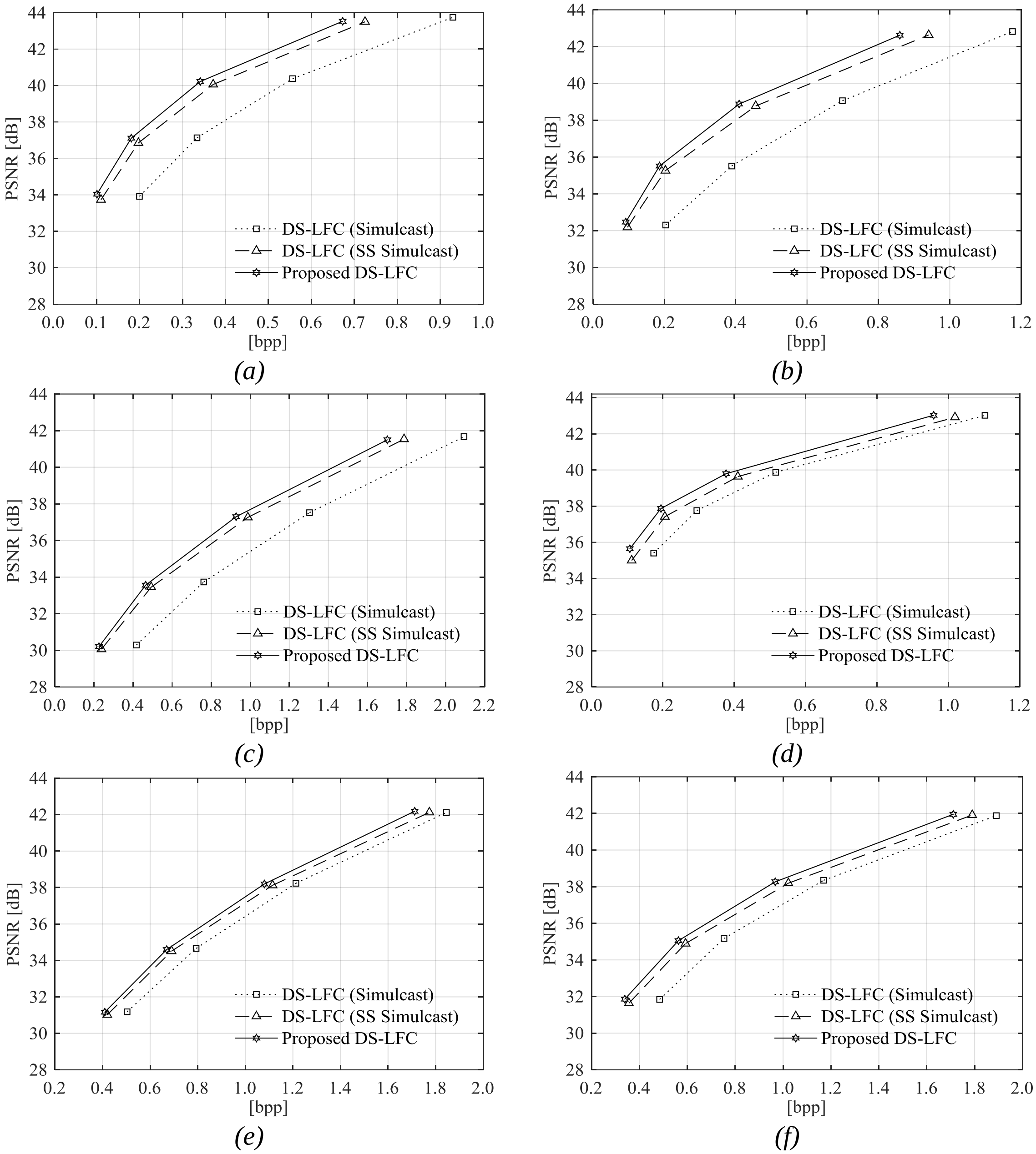
- **Comparison with PVS-Based Approaches** – It can be seen in Table 5.2 that the proposed DS-LFC solution architecture presents expressively better RD performance than the PVS-based arrangement of VIs for both tested configurations (Low Delay P

Table 5.3 RD performance (using BD metrics [210]*) of the proposed DS-LFC codec with respect to the simulcast solutions (QP values 22, 27, 32, and 37). In all solutions, the Basic Rendering algorithm is used for generating content for the lower hierarchical layers*

LF Image	**DS-LFC (Simulcast)**		**DS-LFC (SS Simulcast)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**
Fredo	2.85	-41.32	**0.44**	**-8.52**
Seagull	**3.00**	**-44.56**	0.43	-9.08
Laura	2.59	-33.05	0.35	-5.86
Demichelis Spark (frame 1)	1.14	-29.04	0.26	-7.56
Robot 3D (frame 54)	1.18	-13.02	0.26	-3.12
Plane and Toy(frame 123)	1.53	-20.58	0.34	-5.22
Average	**2.05**	**-33.71**	**0.35**	**-6.56**

and Random Access). The average BD gains of the proposed DS-LFC are 6.26 dB (PSNR) with 68.11 % of bit savings when compared to the VI-based PVS (Low Delay P), and 5.23 dB with 61.03 % of bit savings when compared to VI-based PVS (Random Access). It is worth noting that, for the proposed DS-LFC solution, the 9×1 views in the lower layers are independently encoded as intra frames. Therefore, further bit savings may be expected if inter-view prediction is also allowed for coding the lower layers (as in the PVS-based solutions). Moreover, it is also important to notice that the above conclusions are for LF images captured by using a focused LF camera setup and with different microlens shapes and sizes (see Appendix A). Further comparisons with respect to a larger set of LF camera setups will be considered in future work.

- **Comparison with HEVC (Single Layer)** – As shown in Table 5.2, 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.



(a) (b) (c) (d) (e) (f)

Figure 5.6 DS-LFC RD performance in terms of luma PSNR and bpp (QP values 22, 27, 32, and 37) for the LF test images: (a) Fredo, (b) Seagull, (c) Laura, d) Demichelis Spark (frame 1), (e) Robot 3D (frame 54), and (f) Plane and Toy (frame 123)

- **Comparison with DS-LFC (Simulcast)** – 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 with 33.71 % of bit savings (see Table 5.3 and Figure 5.6). 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 are justified by the efficiency of exploiting the redundancy between the layers (using the proposed IL prediction), as well as the efficiency of exploiting the correlations within the LF enhancement layer (using the SS prediction).

- **Comparison with DS-LFC (SS Simulcast)** – Comparing this solution with the proposed complete DS-LFC solution (see Table 5.3 and Figure 5.6), it can be seen that the proposed DS-LFC has better RD performance with average BD gains of 0.35 dB and 6.56 % of bit savings. As expected, it is shown that improved RD performance can be attained by taking advantage of the redundancy in all domains (local and non-local spatial domain, as well as inter-layer domain).

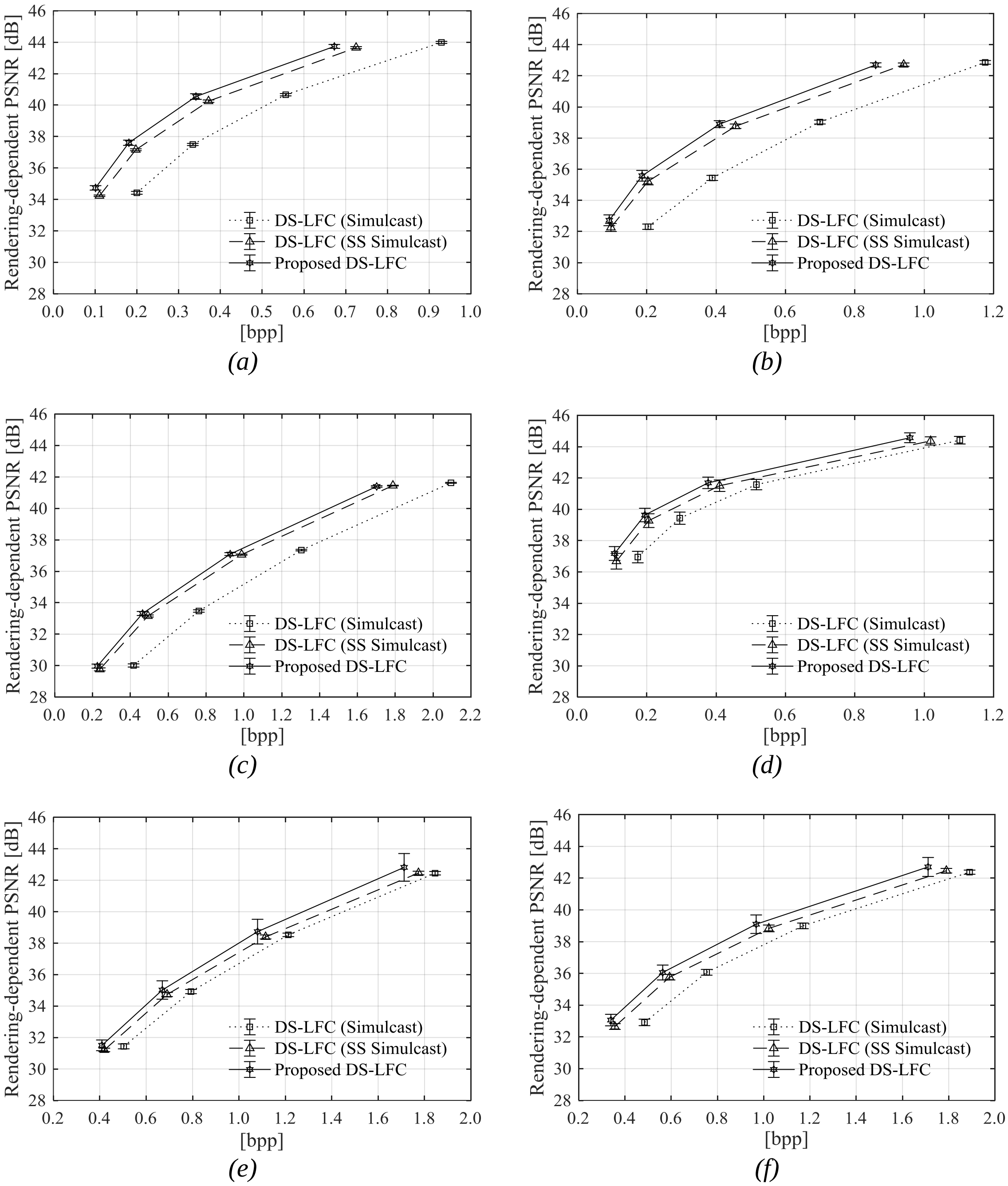
5.5.3 RD Performance for Rendered Views

In order to assess the performance of the proposed scalable coding architecture regarding the quality of rendered views, the RD performance of the proposed DS-LFC solution is here presented in terms of the Rendering-dependent PSNR and SSIM metrics over a set of rendered views (as explained in Section 5.5.1) and compared to the simulcast cases: DS-LFC (Simulcast) and DS-LFC (SS Simulcast). This comparison aims at analyzing the quality improvements achieved at the highest layer by including the proposed IL prediction scheme into the LF enhancement codec. For this reason, other coding architectures (i.e., the PVS-based solutions and the non-scalable HEVC (Single Layer)) are not presented in this analysis. The RD performance is shown in Figure 5.7, in terms of PSNR, and in Figure 5.8, in terms of SSIM.

Moreover, to illustrate the results in terms of visual quality of rendered views, Figure 5.9 shows a portion of a central view rendered from the content in the LF enhancement layer for three different LF test images. For all compared coding solutions, the QP values of the encoder are adjusted to lead to the same bpp value. The results for the remaining LF test images are very similar to the ones presented in Figure 5.9.

For the results in Figures 5.7 to 5.9, the Basic Rendering algorithm was adopted for generating the content in the lower hierarchical layers. From these results, the following conclusions can be drawn:

- **Comparison Between Different Objective Quality Metrics** – Analyzing the results for the two Rendering-dependent metrics (Figures 5.7 and 5.8) and comparing them with the PSNR of the reconstructed LF image (Figure 5.6), it can be seen 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 Rendering-dependent PSNR, the RD gains of the proposed DS-LFC solution go up to 0.76 dB and 14.33 % compared to DS-LFC (SS Simulcast) for LF test image *Fredo* (using the BD metric [210]), being a bit more expressive than for the results presented in Figure 5.6 (up to 0.44 dB and 8.52 %). Regarding the standard deviation values presented in Figures 5.7 and 5.8, a more careful analysis of the PSNR/SSIM 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* and *Plane and Toy* mainly due to the increased vignetting that appears in these images, at the border of each MI (see Figures A.7, A.8b, and A.9).



(a)

(b)

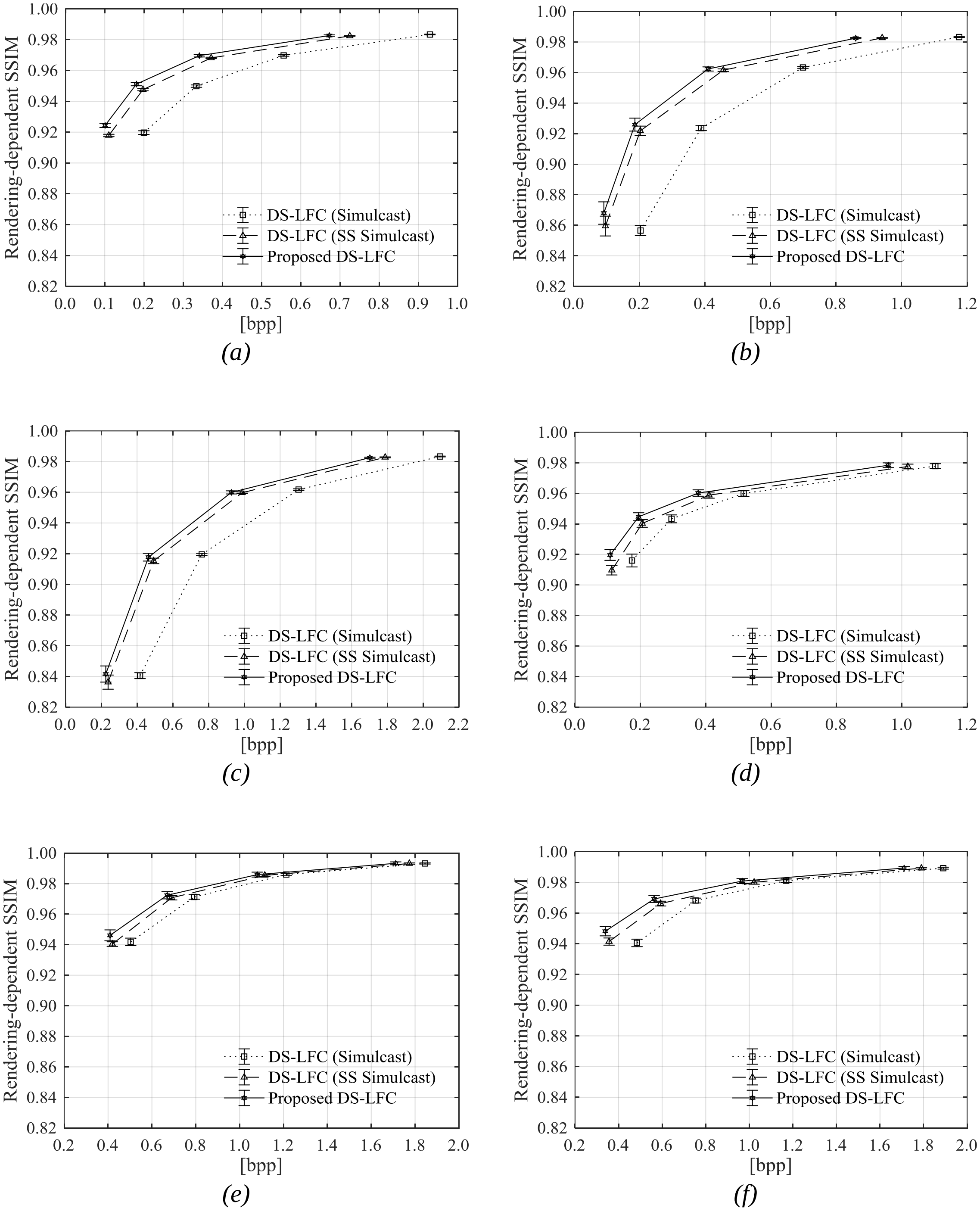
(c)

(d)

(e)

(f)

Figure 5.7 RD performance in terms of PSNR against bpp (QP values 22, 27, 32, and 37) for a set of rendered views from image: (a) Fredo, (b) Seagull, (c) Laura, d) Demichelis Spark (frame 1), (e) Robot 3D (frame 54), and (f) Plane and Toy (frame 123)



(a)

(b)

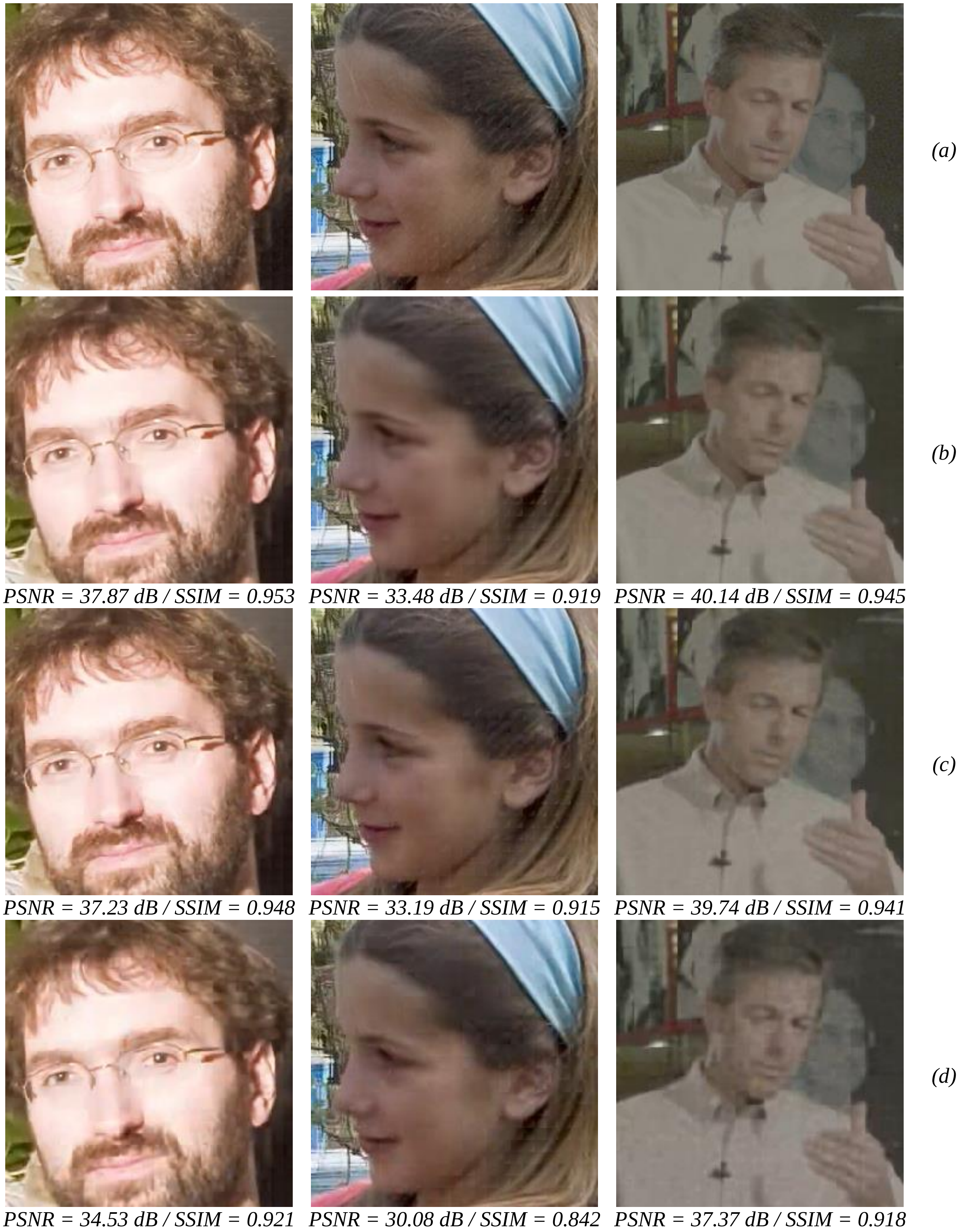
(c)

(d)

(e)

(f)

Figure 5.8 RD performance in terms of SSIM against bpp (QP values 22, 27, 32, and 37) for a set of rendered views from image: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark (frame 1), (e) Robot 3D (frame 54), and (f) Plane and Toy (frame 123)



PSNR = 37.87 dB / SSIM = 0.953 PSNR = 33.48 dB / SSIM = 0.919 PSNR = 40.14 dB / SSIM = 0.945

PSNR = 37.23 dB / SSIM = 0.948 PSNR = 33.19 dB / SSIM = 0.915 PSNR = 39.74 dB / SSIM = 0.941

PSNR = 34.53 dB / SSIM = 0.921 PSNR = 30.08 dB / SSIM = 0.842 PSNR = 37.37 dB / SSIM = 0.918

Figure 5.9 Comparison of a portion of Fredo (left), Laura (middle), and Demichelis Spark (right) when rendering from: (a) original image; (b) encoded frame using DS-LFC (proposed); (c) encoded frame using DS-LFC (SS Simulcast); and (d) encoded frame with DS-LFC (Simulcast). SSIM and PSNR values (Rendering-dependent) are shown for the three test images at around 0.199 bpp for Fredo, 0.417 bpp for Laura, and 0.174 bpp for Demichelis Spark

Table 5.4 RD performance (using BD metrics [210]*) of the proposed DS-LFC codec when using the Weighted Blending algorithm for generating content for the lower hierarchical layers (QP values 22, 27, 32, and 37)*

LF Image	**DS-LFC (Simulcast)**		**DS-LFC (SS Simulcast)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**
Fredo	2.97	-41.79	0.43	-8.05
Seagull	2.96	-47.38	0.42	-9.51
Laura	2.23	-34.89	0.28	-5.86
Demichelis Spark (frame 1)	1.31	-31.60	0.28	-7.60
Robot 3D (frame 54)	1.09	-12.74	0.18	-2.28
Plane and Toy(frame 123)	1.51	-21.43	0.31	-4.99
Average	**2.01**	**-31.64**	**0.32**	**-6.38**

- **Comparison with Visual Quality Inspection of Rendered Views** –. From the results in Figure 5.9, it can be seen that the proposed DS-LFC solution also improves the visual quality of views rendered from the LF hierarchical layer. Artifacts when coding with DS-LFC (Simulcast) are more evident, such as in the eyes of *Laura*, and all over the face of *Demichelis Spark*. Moreover, it can be observed that these results are also consistent to the results obtained with the Rendering-dependent quality metrics. Improvements are also noticeable when compared to the DS-LFC (SS Simulcast) solution, for instance, in the man's face pictured behind the presenter in *Demichelis Spark*.

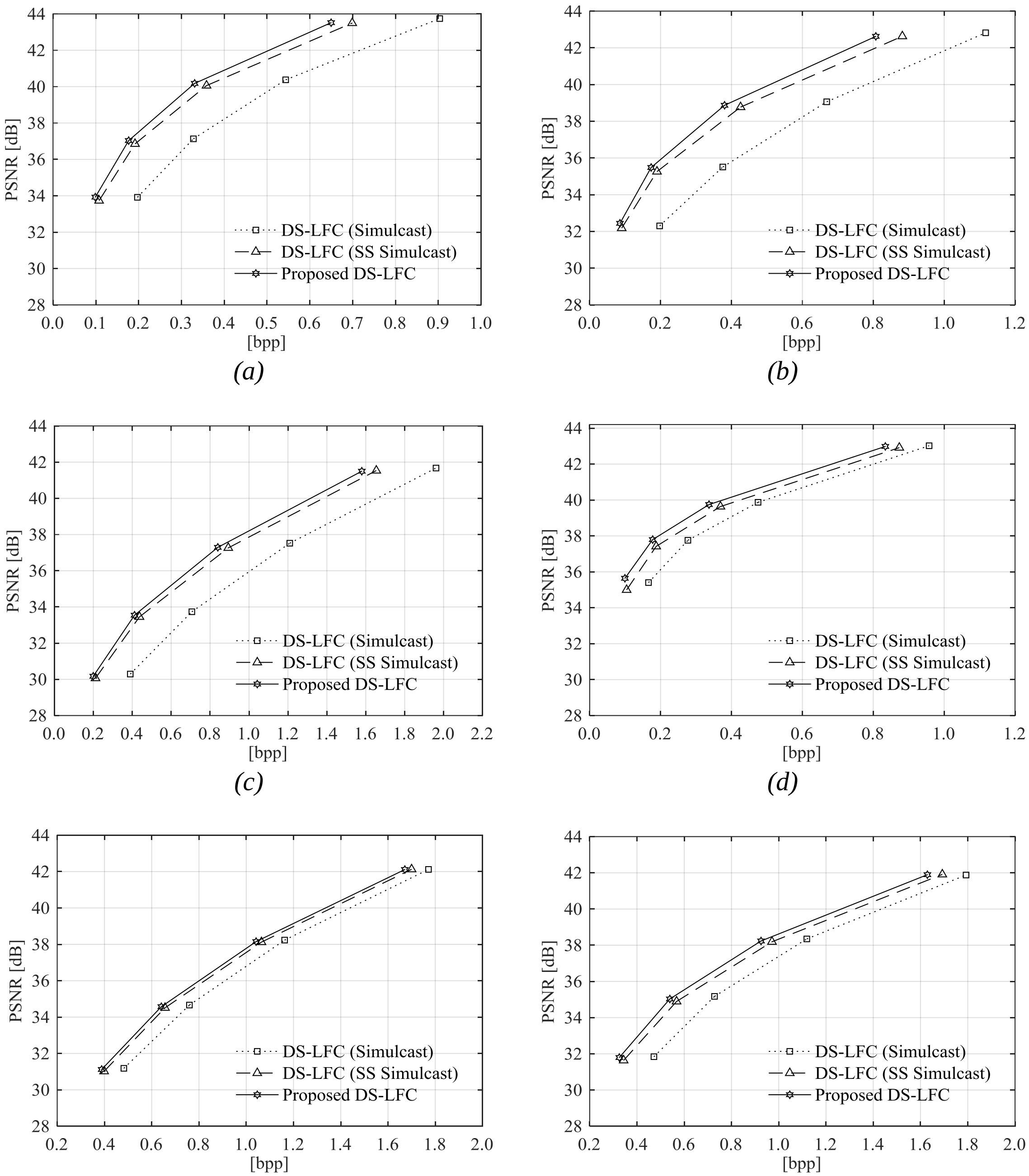
5.5.4 RD Performance for Different Acquisition Parameters

In order to further discuss the usability of the proposed scalable coding architecture, the performance of the proposed DS-LFC solution is here analyzed in terms of different acquisition parameters for generating the content in the 2D and stereo or multiview layers (i.e., base and first enhancement layers).

Therefore, this analysis considers the results in terms of:

- **Different Rendering Algorithms** – This analysis compares the RD results between using the Basic Rendering [10] algorithm (Table 5.3, and Figures 5.6 to 5.8) and the Weighted Blending algorithm [10] to generate the views for the lower layers. For this, the BD results using the Weighted Blending algorithm are shown in Table 5.4, and the RD performance is depicted in Figure 5.10 (in terms of the PSNR for the entire content in the LF enhancement layer), Figure 5.11 (in terms of Rendering-dependent PSNR metric), and Figure 5.12 (in terms of Rendering-dependent SSIM metric). It can be seen that, there are no significant differences when comparing the BD results in Tables 5.3 and 5.4 and the RD results in Figures 5.6 and 5.11. On the other hand, regarding the Rendering-dependent metrics, it can be observed that when using the Weighted

Blending algorithm (see Figures 5.11 and 5.12), much higher SSIM values are observed than when using the Basic Rendering algorithm (see Figures 5.7 and 5.8). This can be explained by the high level of blur which is introduced by the weighted average in the Weighted Blending algorithm [10]. Hence, this weighted average works as a filter masking some of the coding artifacts in the rendered views.

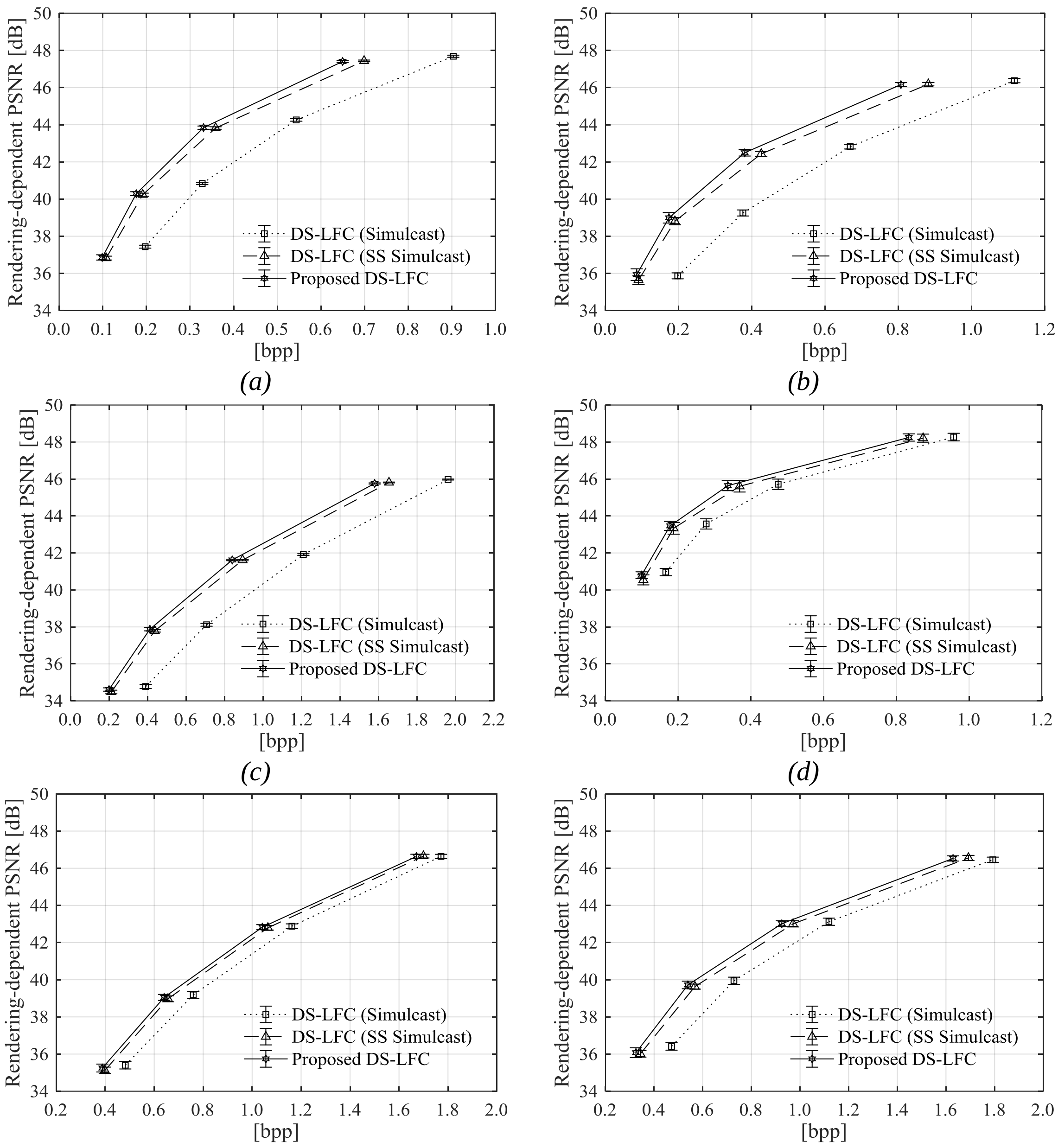


(a) *(b)*

(c) *(d)*

(e) *(f)*

Figure 5.10 RD performance when using Weighted Blending as the rendering algorithm for: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark (frame 1), (e) Robot 3D (frame 54), and (f) Plane and Toy (frame 123)



(a) *(b)* *(c)* *(d)* *(e)* *(f)*

Figure 5.11 RD performance in terms of PSNR against bpp (QP values 22, 27, 32, and 37) for a set of views rendered using Weighted Blending algorithms from image: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark (frame 1), (e) Robot 3D (frame 54), and (f) Plane and Toy (frame 123)

- **Different Patch Sizes** – This analysis compares the results when using different patch sizes for generating the content in the lower layers (for the same rendering algorithm and patch positions). To illustrate this, Table 5.5 presents some prediction usage statistics and BD results for the test image *Fredo* so as to understand how the patch size parameter affects the accuracy of the IL prediction. More specifically, Table 5.5 shows the coding statistics and the BD performance of the proposed DS-LFC for two different patch sizes: i) patch size 10 corresponds to the case where the *Fredo*'s face, most of his body, and portions of the composing scene is in focus (see Figure A.1); and ii) patch size 15 corresponds to the case where only *Fredo*'s hand and the object he is holding are in focus (see Figure A.1). From the presented results (see Table 5.5), is can be observed that the patch size 10, where a larger portion of the scene is in focus, presented a larger percentage of IL prediction usage (6.76 %) associated with a better RD coding performance (with BD gains of 2.83 dB and 40.70 % of bit savings when compared to the DS-LFC (Simulcast) solution). This knowledge may be particularly useful in a scene where there is an interest in varying the patch sizes from one time instant to another (e.g., for the *Plane and Toy* sequence in Figure A.8).

- **Different Number of 2D View Images in The Lower Layers** – This analysis compares the results when different numbers of 2D views are generated to compose the content in the lower layers (but using the same patch sizes, as seen in Table 5.1, and rendering with the Basic Rendering algorithm). For this, Table 5.6 shows the coding statistics and the BD performance for *Fredo* when two different sets of views are acquired for the lower hierarchical layers: i) 5×1 views are rendered; and ii) 9×1 views are rendered. As expected, the results in Table 5.6 show that the more 2D views are considered in the lower layer, the larger the percentage of usage of the Inter layer prediction will be used and, consequently, the better the RD performance of the proposed DS-LFC codec will be.

5.6 Final Remarks

This chapter proposed a three-layer scalable coding architecture for LF content so as to provide backward compatibility with legacy 2D and 3D displays. Based on this architecture, an efficient prediction scheme was also proposed for improving the RD performance in the LF enhancement layer. Notably, an IL prediction scheme was proposed which built a LF prediction reference from the reconstructed multiview content in the lower layers. Moreover, the SS compensated prediction was also proposed to be added to the LF enhancement layer coder so as to take advantage of the redundancy in all domains (local and non-local spatial domain, and inter-layer domain).

Experimental results confirmed the advantage of the proposed scalable coding architecture when compared to various benchmark solutions. Notably, the proposed DS-LFC solution outperformed with significant gains the simulcast cases (BD gains of up to 3.00 dB and 44.46 % of bit savings), and it was shown that it is possible to provide display scalability and still present better RD performance than a non-scalable solution using HEVC Still Picture Profile (with BD gains of up to 2.40 dB and 37.90 % of bit savings).

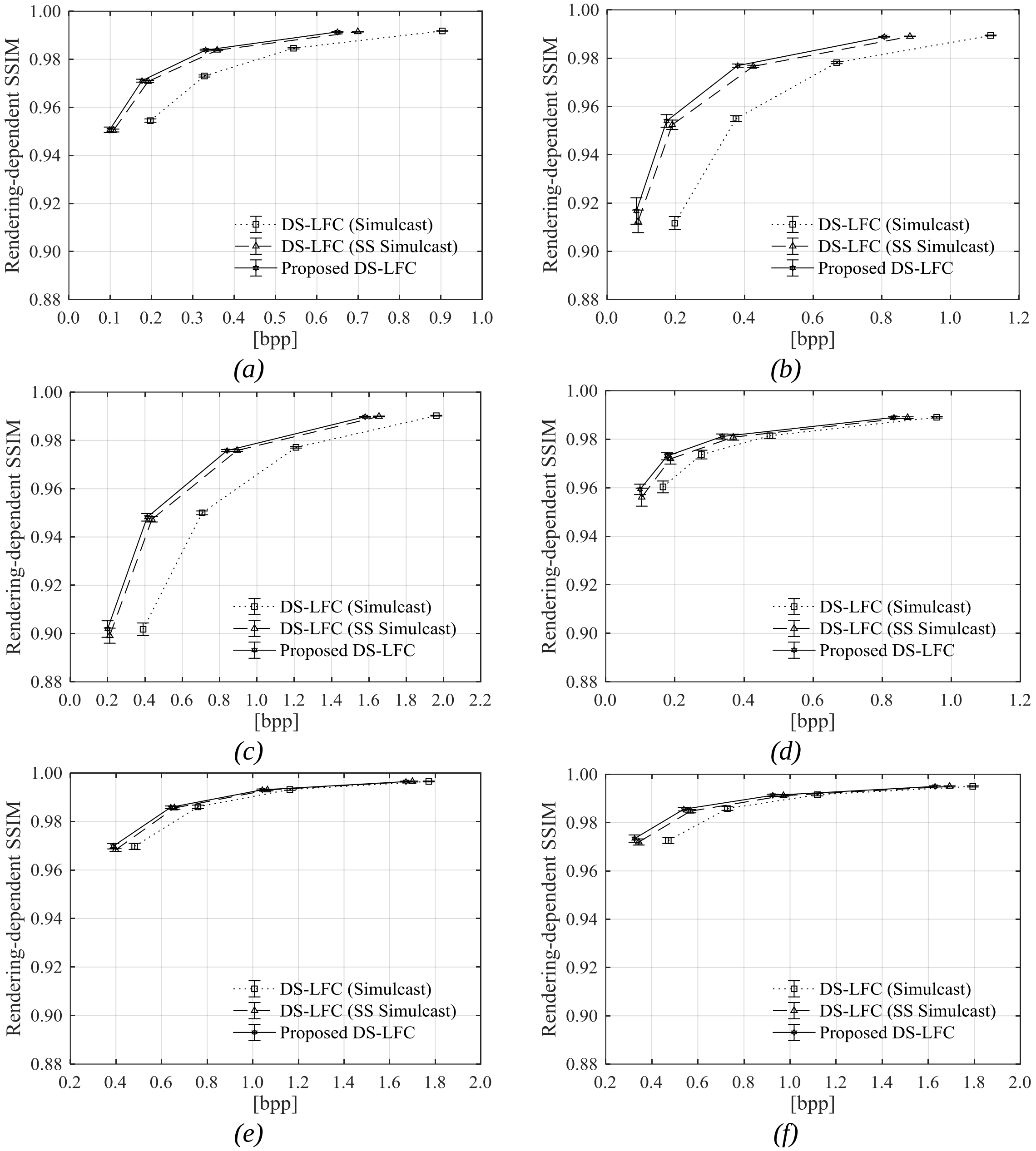


Figure 5.12 RD performance in terms of SSIM against bpp (QP values 22, 27, 32, and 37) for a set of views rendered using Weighted Blending algorithms from image: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark (frame 1), (e) Robot 3D (frame 54), and (f) Plane and Toy (frame 123)

Table 5.5 Prediction mode statistics, BD-PSNR, and BD-BR results for test image Fredo using different patch sizes

Patch Size	**Prediction Mode Statistics**					**Proposed DS-LFC vs DS-LFC (Simulcast)**		**Proposed DS-LFC vs DS-LFC (SS Simulcast)**	
	Intra	**Skip**	**SS**	**IL**	**Bi-Pred.**	**BD-PSNR**	**BD-BR**	**BD-PSNR**	**BD-BR**
10	26.13 %	13.07 %	21.22 %	6.76 %	32.82 %	2.83 dB	-40.70 %	0.63 dB	-11.56 %
15	30.25 %	14.37 %	23.31 %	3.59 %	28.49 %	2.27 dB	-33.49 %	0.46 dB	-8.36 %

Table 5.6 Prediction mode statistics, BD-PSNR, and BD-BR results for test image Fredo using different number of views in the lower hierarchical layers

Num. Views	**Prediction Mode Statistics**					**Proposed DS-LFC vs DS-LFC (Simulcast)**		**Proposed DS-LFC vs DS-LFC (SS Simulcast)**	
	Intra	**Skip**	**SS**	**IL**	**Bi-Pred.**	**BD-PSNR**	**BD-BR**	**BD-PSNR**	**BD-BR**
5	30.93 %	14.07 %	22.61 %	2.54 %	29.85 %	2.57 dB	-37.12 %	0.44 dB	-8.09 %
9	26.13 %	13.07 %	21.22 %	6.76 %	32.82 %	2.83 dB	-40.70 %	0.63 dB	-11.56 %

Chapter 6

Light Field Coding with Field of View Scalability for Flexible Interaction

It was seen in the previous chapter that, as imaging technologies move towards richer representations, novel data coding solutions are essential to gradually support the new applications and functionalities that arise [31]. Therefore, the scalable coding solution proposed in Chapter 5 focused on building the bridge between legacy 2D/3D and future LF imaging applications. Differently, this chapter goes a step further and focuses on future LF imaging applications, as well as on how to hierarchically organize the huge amount of LF information for supporting LF applications with flexible end-user interaction.

Among the exciting new interactive functionalities that future LF imaging applications support are the possibility for post-production refocusing, changing depth of field, and changing viewing perspective. This means that, for instance, the end-user can receive a captured LF content and interactively adjust the plane of focus and depth of field in 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, a scalable 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 (as illustrated in Figure 6.1): 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 aforementioned requirement (see Figure 6.1), this chapter proposes a new scalability concept, named FOV scalability, as well as a novel Field Of View Scalable Light Field Coding (FOVS-LFC) solution. Taking advantage of the 4D radiance distribution, the FOV scalability progressively supports richer forms of the same content by hierarchically organizing the angular information of the captured LF. 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

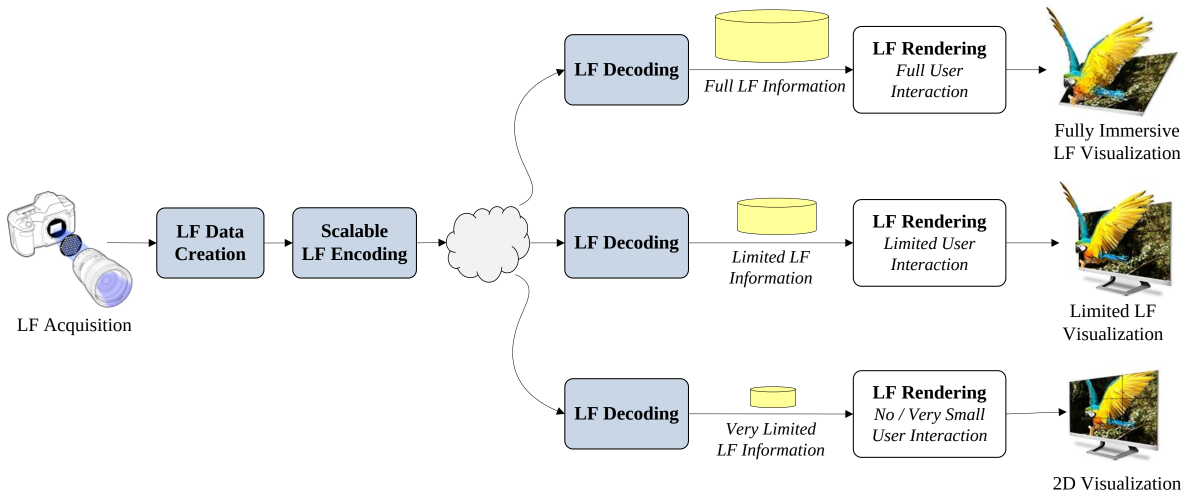


Figure 6.1 Requirement for supporting flexible interactive functionalities in the proposed FOV scalable LF coding solution

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 (see Figure 6.1) 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 subsets of the bitstream (that fit in the available bitrate).

In addition to the FOVS-LFC architecture, two novel IL prediction schemes are also proposed in this chapter to be used as alternative prediction modes to efficiently encode the LF data in each enhancement layer. These are:

- **Exemplar-Based Direct IL Prediction** – In this case, a direct prediction scheme is used, as proposed in Section 6.3.1, where exemplar texture samples [215] from lower layers are used for implicitly estimating a good prediction block. This way, no further information about the used predicton block needs to be transmitted to the decoder side.
- **IL Prediction with Patch-Based ILR Picture Construction** – In this case, an IL compensated prediction scheme is adopted, in which an ILR picture that is constructed, as presented in Section 6.3.2, relying on a patch-based optimization algorithm for texture synthesis is used.

In a nutshell, the FOVS-LFC solution proposed in this chapter provides: i) richer and flexible rendering capabilities (such as refocusing, changing perspective and depth of field changing); ii) high compression efficiency; iii) high quality of rendered views in all layers; as well as, iv) backward compatibility with the current state-of-the-art HEVC standard [2].

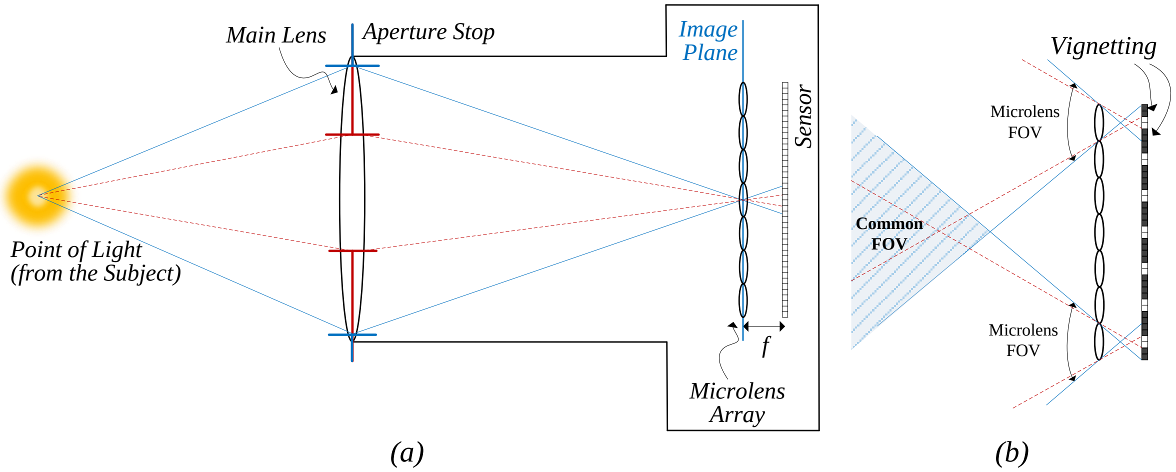


Figure 6.2 Microlens FOV in the LF camera: (a) Influence of the main lens aperture size in the microlens FOV (note that this is an illustrative example where the geometrical optics is not proportionally represented); (b) The common FOV, i.e. the area where the FOV of all microlenses overlaps, can be used as a measure of overall angular information in the LF content

The remainder of this chapter is organized as follows. Section 6.1 presents the concept of FOV scalability, while Section 6.2 describes the FOVS-LFC coding architecture. Section 6.3 describes the novel IL coding tools for improving the compression performance. Section 6.4 presents the test conditions and experimental results; and, finally, Section 6.5 concludes the chapter.

6.1 Proposal for a Data Organization with FOV Scalability

In this section, the proposed scalable data representation is presented by directly looking at the distribution of the 4D light field [77] inside the LF camera so as to better illustrate the concept of FOV scalability. For this, Sections 6.1.1 and 6.1.2 firstly overview the concept of microlens FOV and radiance in an LF camera. Then, the FOV scalability concept is presented in Section 6.1.3, and the proposed coding architecture based on the FOV scalability concept is theoretically compared to other scalable LF coding architectures in Section 6.1.4.

6.1.1 Microlens 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; or, in other words, the area over which the sensor ‘sees’ through the lens. In a conventional 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, Figure 6.2a depicts the traditional 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 microlens array (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 Figure 6.2a), it is possible to see that the main lens aperture (or more specifically, the F-number[3] of the main lens) needs to be matched to the F-number of the microlens array to guarantee that MIs receive homogeneous illumination in its entire area, as seen in the blue line case (Figure 6.2a). Otherwise, in the case of the dashed red line (Figure 6.2a), pronounced vignetting (with the shape of the main lens aperture) will be visible in each MI. An illustrative example is given in Figures A.7-A.9.

As depicted in Figure 6.2b, 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. Therefore, if there is MI vignetting (see dashed red lines in Figure 6.2b), the microlens FOV will be further restricted and, consequently, the angular information in the captured LF image will be narrowed.

6.1.2 Radiance Distribution in LF Cameras

The radiance is basically a function defined in 4D space, in which a light ray is described by the position (x, y) at which it intersects a plane perpendicular to the optical axis and by the corresponding ray slope (θ, φ). To simplify the visualization of this 4D function (coordinates x, y, θ and φ), the flat Cartesian ray-space diagram [77, 83] shown in Figure 6.3a is used, where only two dimensions – in this case, x and θ – are considered. For more details about LF parameterization, please refer to [77, 83].

Therefore, the ray-space diagram in Figure 6.3b illustrates the radiance at the microlens plane for the camera setup in Figure 6.2a. Briefly, the radiance coming from the captured scene is refracted through the main lens and then split into each microlens in the array. This can be seen in Figure 6.3b, 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 Figure 6.3b, corresponds to the tiny bundle of rays that is integrated into a single pixel of the MI.

Furthermore, the radiance distribution of the traditional LF camera (Figure 6.3b) can be generalized to the focused LF camera setup [10] (see Figure 2.5). As discussed in Chapter 2, in this case, the main lens and the microlenses are focused in an image plane in front (or behind) of the microlens array plane. As a result, the main lens forms a relay system with each microlens. Therefore, as shown in [83], each MI will then capture what corresponds to a slanted stripe of the radiance at the image plane, as depicted by the ray-space diagram in Figure 6.3c. As a result, this configuration allows an effective increase in spatial resolution at

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

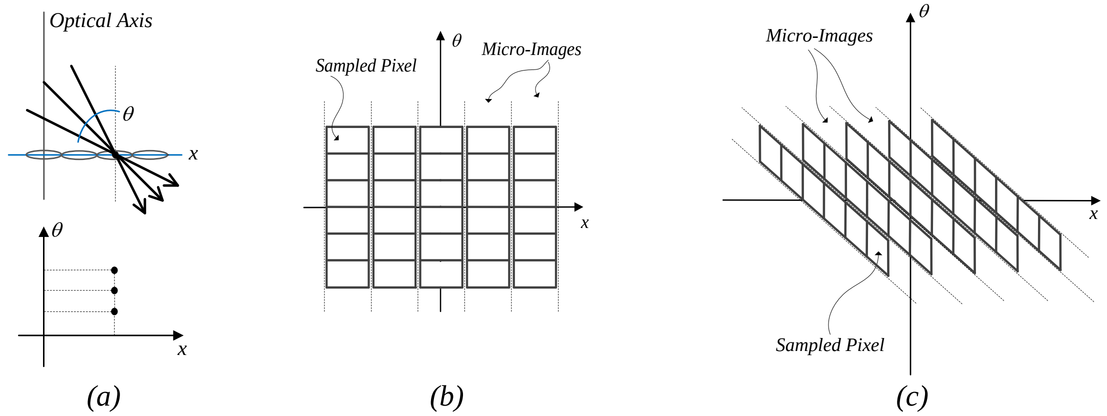


Figure 6.3 Spatio-angular parameterization of the radiance in a LF camera: (a) A single ray of light is described by the position it intersects the plane x and its slope θ. Each possible ray in the ray diagram (top) corresponds to a different point in the Cartesian ray-space diagram (bottom); (b) Sampling the radiance at the image plane for the traditional LF camera; and (c) Sampling the radiance at the image plane for the focused LF camera

the price of a decrease in directional resolution. The generalized focused LF camera setup will be considered hereinafter.

6.1.3 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 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).

As was shown in the previous section, each pixel underneath its corresponding microlens gathers light information from a given direction. Therefore, it is possible to split the overall angular information by properly selecting subsets of pixels from each MI. This concept is illustrated in Figure 6.4 for a hypothetical case in which three subsets of pixels are sampled from each MI (see Figure 6.4a). Therefore, the angular information is split into three layers as seen in Figure 6.4b. In each lower layer (from top to bottom in Figure 6.4b), the microlenses FOV will be further restricted and, consequently, the angular information of the system will be narrowed (as depicted in Figure 6.4a).

Since the central angular direction is usually the most important (it is usually the perspective the shooter will point to when acquiring the LF image), the angular information is chosen to grow from the central to the border samples in each MI. For instance, Figure 6.5 shows the selection of three subsets of pixels with different angular information from each MI to build a three-layer FOV scalable LF representation. For the base layer (Figure 6.5a), a set with only the central angular information is gathered. For the enhancement layers 1 and 2 (respectively, Figure 6.5b and Figure 6.5c), the samples progressively contain wider angular information

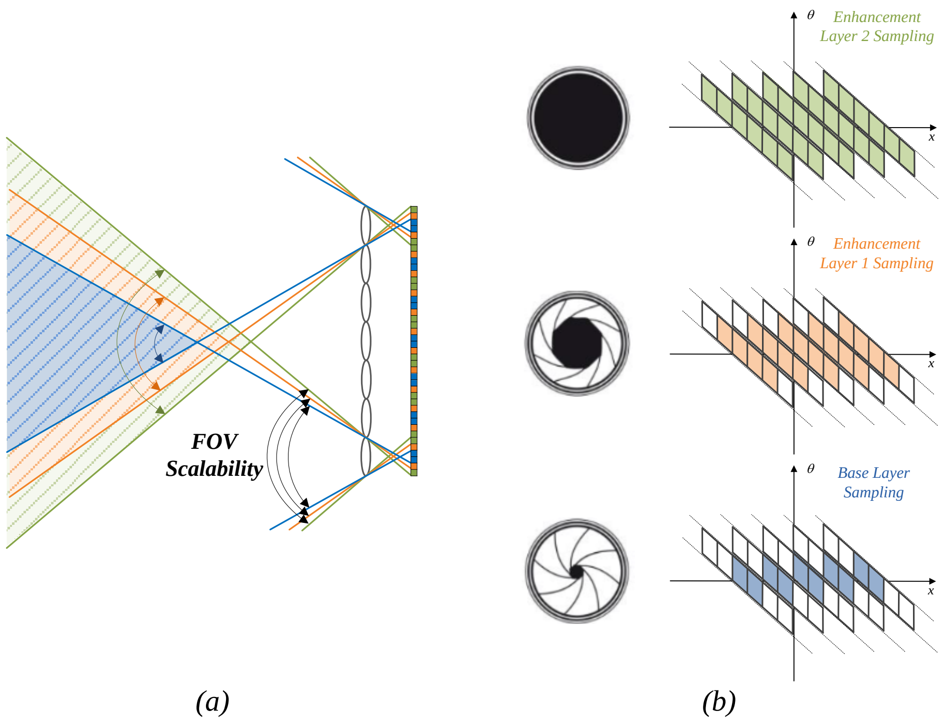


Figure 6.4 The concept of FOV scalability in a LF system: (a) Ray tracing diagram showing that three hierarchical layers of FOV scalability can be sampled by properly selecting three subsets of pixels (with different colors) from each MI, corresponding to a progressively larger amount of angular information; and (b) Corresponding ray-space diagram showing the angular information in each hierarchical layer

(from the center to the borders). A consequence of the increased angular information is that the resolution of the LF image in each layer will also increase from one layer to the next higher layer.

With this representation, it is then possible to define new levels of scalability 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 to interact with.
- **Changing Focus (Refocusing)** – As discussed in Chapter 2, refocusing can be seen as virtually translating the image plane of the LF camera (see Figure 6.2a) from the microlens plane to another one 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.

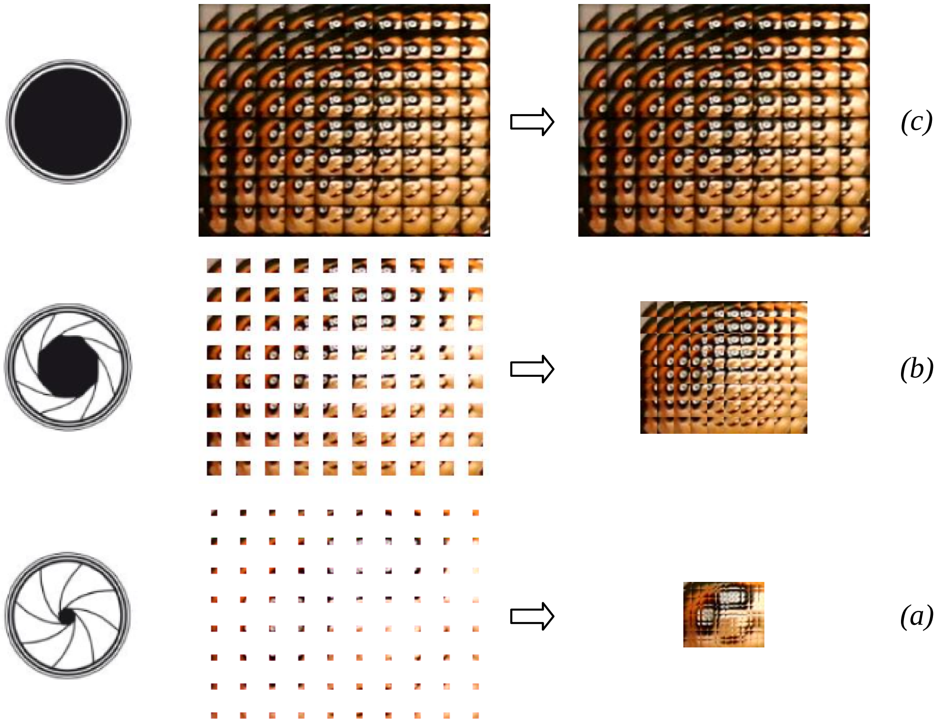


Figure 6.5 Gathering information for three hierarchical layers to support FOV scalability in a portion of the LF image Plane and Toy (frame 123) as shown in Figure A.8b: (a) base layer; (b) enhancement layer 1; and (c) enhancement layer 2. From the base layer (a) to the last Enhancement layer (c), the FOV is wider and, consequently, the angular information progressively grows as well

- **Varying Depth of Field** – Increasing or decreasing the depth of field in LF images simply means to define greater 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 the depth of field is that can be selected during the user's interaction.

The great advantage of the proposed FOV scalable format is the increased flexibility it gives the content creator for 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. For instance, the author decides 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.

Notice that, although the angular information is limited in each MI, it is still possible to derive disparity or ray-space information, and to reconstruct texture information that is still not available in a lower hierarchical layer at the receiver side (although this is out of the scope of this chapter).

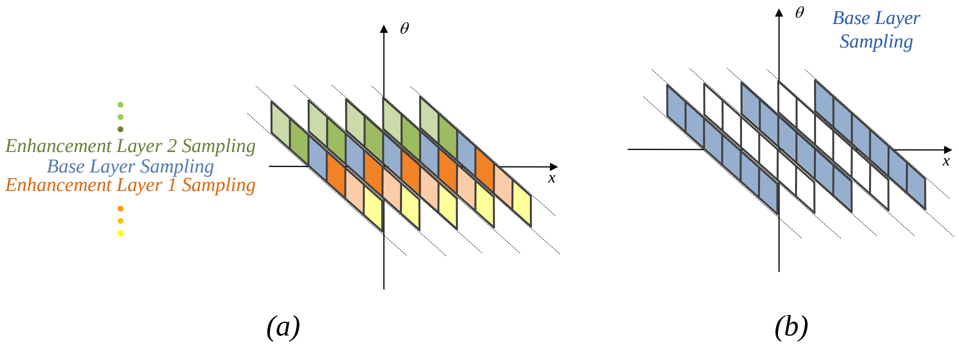


Figure 6.6 Ray-space diagram showing the gathering of information in each hierarchical layer in alternative scalable solutions: (a) PVS- and Multiview-based approaches; and (b) Subsampled grid of MIs

6.1.4 Comparison to Other Scalable LF Coding Architectures

As an illustrative example, the radiance distribution by using other scalable LF representations proposed in the literature is presented in Figure 6.6, notably, when using: i) a PVS- or multiview-based representation [92, 175, 179, 181, 182] (see Figure 6.6a); and ii) a subsampled grid of MIs [204–206].

Regarding the PVS- or multiview-based representation, each layer is selected as a different VI. In the specific case of LF images captured with the focused LF camera setup (shown in Figure 6.6), the texture resampling from MIs to the viewpoint images usually results in very low resolution images with significant aliasing artifacts [86]. Moreover, due to the small resolution of these VIs, the overhead to encode all of them may restrict the usability of this approach.

Regarding the scalable approach based on a subsampled grid of MIs, although the compression ratio is improved by discarding many MIs, the quality of views that are rendered in the lower layers is affected by the quality of the derived disparity/depth information.

Therefore, although the same interactive functionalities can be achieved with these two scalable coding architectures, with respect to compression efficiency and quality of rendered views, the proposed FOV scalable representation can be seen as a good compromise between them both since it is less complex and produces much less overhead than the multiview approach, and supports many different rendering algorithms in which more accurate views are likely to be rendered at lower layers when compared to the subsampled grid of MIs.

6.2 Proposal for a FOV Scalable LF Coding Architecture

The coding architecture proposed in this chapter to provide FOV scalability is built upon a predictive and multi-layered scalable approach, as depicted in Figure 6.7. The base layer

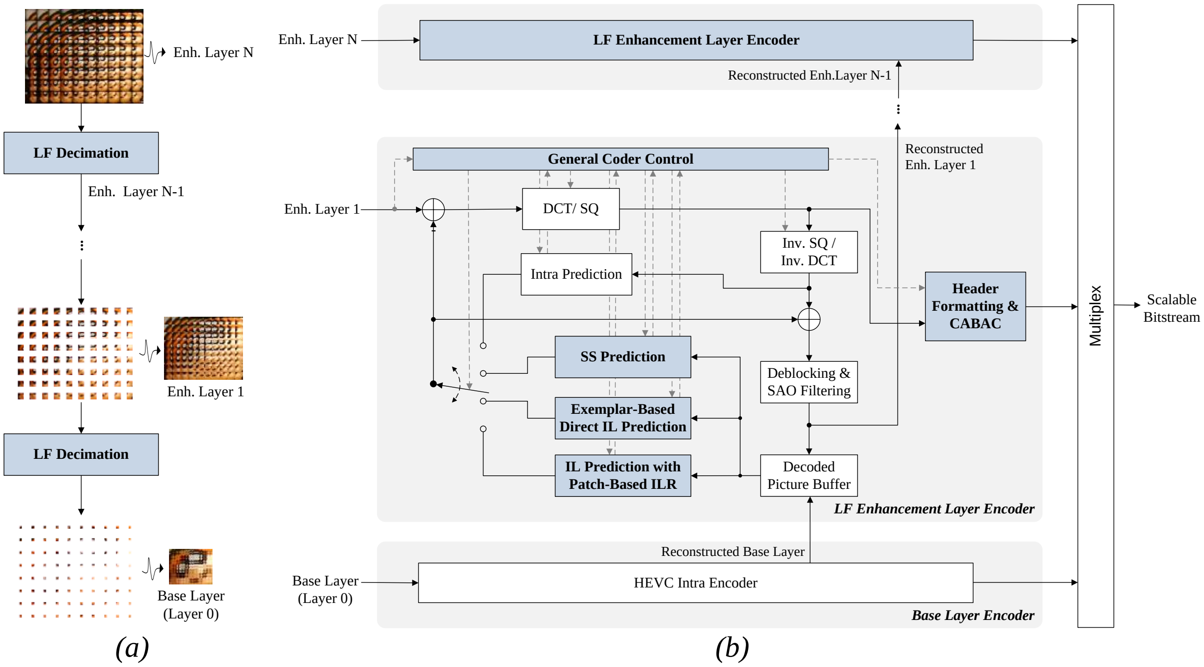


Figure 6.7 Block diagram of the proposed scalable codec: (a) The LF decimation process to generate content for each hierarchical layer; (b) Proposed encoder architecture in which one or more enhancement layers (from 1 to N) are encoded with the proposed FOVS-LFC encoder (novel and modified blocks are highlighted in shaded blue boxes)

contains a sub-sampled portion of the LF raw 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). 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 frames are used for coding the higher layers. Following the base layer, one or more enhancement layers (enhancement layers 1 to N in Figure 6.7) are defined to represent the necessary information to obtain more immersive LF visualization. Each higher enhancement layer 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, as illustrated in Figure 6.7.

The basic blocks (see Figure 6.7) of the proposed FOVS-LFC codec are explained in the following.

6.2.1 LF Decimation

As illustrated in Figure 6.7, the LF raw data is firstly decimated into several layers, where higher layers contain LF content with wider FOV and wider angular information (and consequently LF content with larger resolution).

During a creative authoring process, 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 discussed in Section 6.1.3, 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.

6.2.2 Exemplar-Based Direct IL Prediction

This new prediction mode aims at exploiting the redundancy between adjacent layers to find a prediction block and, then, to implicitly derive an IL prediction for encoding the current block in an enhancement layer picture. As a result, no vector needs to be transmitted and the decoder can simply use the same process for inferring the vectors to carry out the compensated prediction using the decoded residual samples. In order to distinguish the exemplar-based direct IL prediction from the conventional HEVC merge mode [2], an index is transmitted (together with the coded residual information). The process to derive the implicit IL vector is presented in Section 6.3.1.

6.2.3 IL Prediction with Patch-Based ILR

This prediction mode can be used to further improve the enhancement layers coding efficiency by removing redundancy between adjacent layers. In this sense, an enhanced ILR picture is constructed, which can be used as a new reference picture when encoding the current enhancement layer. If IL prediction mode is used, the residual information and an IL vector are coded and transmitted to the decoder. To construct the enhanced ILR picture, a novel coding tool, referred to as patch-based ILR picture construction, is proposed in Section 6.3.2.

6.2.4 SS Prediction

Since the proposed data organization still presents high redundancy between adjacent MIs (or decimated MI texture samples), the SS prediction, previously proposed in Chapter 4, can be also used as an alternative prediction to exploit the existing redundancy and to improve coding efficiency within each enhancement layer. As a result, the residual information and SS vector(s) are coded and sent to the decoder.

6.2.5 Intra Prediction

HEVC intra prediction modes are also available as an alternative prediction when selecting the most efficient mode for each coding block. The decision to choose between intra, SS, exemplar-based direct and IL prediction is made in an RDO manner, as in conventional HEVC.

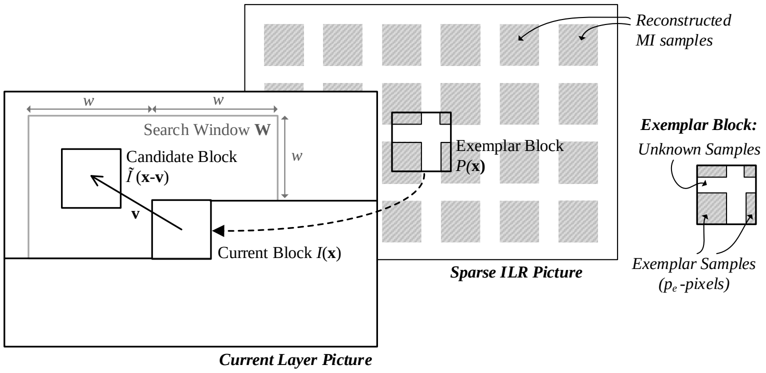


Figure 6.8 Exemplar-based direct IL prediction. The best match between the co-located block and a candidate block (within the causal search Window **W** *in the current picture) allows finding the IL vector of the current block*

6.2.6 DCT and Scalar Quantization (SQ)

Residual information is transformed and quantized using the standard HEVC DCT and SQ to achieve further compression in each layer.

6.2.7 Header Formatting & CABAC

Additional high level syntax elements are carried through the HEVC bitstream to support this new type of scalability. These are basically acquisition information (e.g., MI resolution and additional decimation information) and dependency information (for signaling the use of novel reference pictures). Finally, residual and prediction mode signaling data are entropy coded using CABAC.

6.3 Proposal for Novel IL Coding Tools

To achieve a high compression efficiency, the proposed FOVS-LFC solution relies on the two IL coding tools that are proposed in this section: i) exemplar-based direct prediction; and ii) patch-based ILR picture construction.

6.3.1 Exemplar-Based Direct IL Prediction

When encoding the current block in an enhancement layer picture, this new prediction mode makes it possible to implicitly derive an IL prediction block based on the texture information from a reference layer, as illustrated in Figure 6.8. This process to derive the IL prediction can be divided in the following two steps.

6.3.1.1 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 Figure 6.8. This sparse picture is hereinafter referred to as sparse ILR picture.

The output of this step is the exemplar block, $P(\mathbf{x})$, with same size and co-located position to the current block, $I(\mathbf{x})$, which is derived from the sparse IL reference picture (Figure 6.8).

6.3.1.2 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 [198]) for estimating the 'best' prediction 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 (Figure 6.8). 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 (Figure 6.8) 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 (Figure 6.8). This candidate predictor block is displaced from $I(\mathbf{x})$ by the vector $\mathbf{v}$ (Figure 6.8). Since P contains $(p - p_e)$ unknown samples, it can be modeled as P =A Ĩ, where **A** is a binary mask in which only the corresponding known p_e sample positions are non-zero. Thus, **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 **A** is known a priori, the 'best' candidate predictor block can be simply found by solving the matching algorithm in (6.1).

$$\min_{\mathbf{v},\tilde{I}(\mathbf{x}-\mathbf{v})\subset\mathbf{W}} \left\| P(\mathbf{x}) - \mathbf{A}\tilde{I}(\mathbf{x}-\mathbf{v}) \right\|_1 \tag{6.1}$$

To keep the complexity low, the predictor block is searched inside a limited search window, **W**, as depicted in Figure 6.8 (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 (6.1).

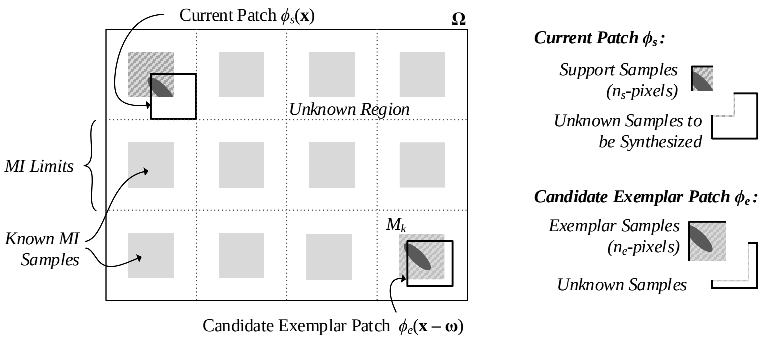

Current Patch $\phi_s(\mathbf{x})$
$\boldsymbol{\Omega}$
Unknown Region
MI Limits
Known MI Samples
M_k
Candidate Exemplar Patch $\phi_e(\mathbf{x}-\boldsymbol{\omega})$
Current Patch ϕ_s:
Support Samples (n_s-pixels)
Unknown Samples to be Synthesized
Candidate Exemplar Patch ϕ_e:
Exemplar Samples (n_e-pixels)
Unknown Samples

Figure 6.9 Patch-based ILR picture construction: For each patch ϕ_P in the sparse ILR picture, the best candidate patch, ϕ_c^{best} is derived by solving the optimization problem in (6.2)

6.3.2 Patch-Based ILR Picture Construction for IL Prediction

In order to further improve the coding efficiency, an enhanced ILR picture is built, which is used by the IL prediction (as shown in Figure 6.7) when encoding the current enhancement layer. This section describes the patch-based process for building this enhanced ILR picture.

6.3.2.1 Input Information

Similarly to the exemplar-based direct IL prediction, the input information for this process is the coded and reconstructed frame from a reference layer. Hence, the corresponding sparse ILR picture (see Figure 6.9) is derived, which is used as the input information for synthesizing the enhanced ILR picture in the following.

6.3.2.2 Problem Formulation

The basic idea for constructing the enhanced ILR picture is to find a good estimation to fill in the unknown data in the sparse ILR 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.

Firstly, the unknown pixels are initially set to zero. Moreover, this sparse ILR picture is divided into n-pixel non-overlapping patches, ϕ_s, to apply the texture synthesis algorithm (see Figure 6.9). 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 (Figure 6.9). 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 ILR 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 Figure 6.9. 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, **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 (6.2),

$$\min_{\phi_e(\mathbf{x}-\boldsymbol{\omega})\subset\boldsymbol{\Omega},\mathbf{A}} \left\|\mathbf{B}\left(\phi_s(\mathbf{x})-\phi_e(\mathbf{x}-\boldsymbol{\omega})\right)\right\|_1+\lambda\times\left\|diag\left(\mathbf{I}_n-\mathbf{B}\right)\right\|_0 \tag{6.2}$$

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

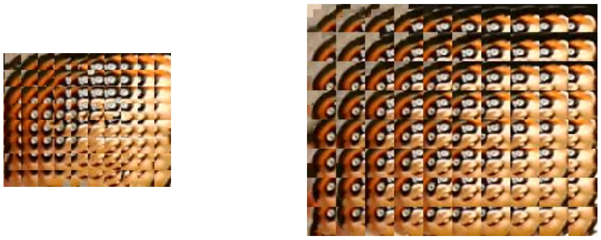
The problem in (6.2) comprises a data-fitting term and a sparseness prior function, respectively. The former term tries to find the best match within the region where ϕ_s and ϕ_e overlap, while the latter term penalizes candidate patches whose n_e-pixel region is too small. In addition to this, since 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 (6.2) 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 (6.2), 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 in Section 6.4, the λ value is selected empirically. The patch size is selected to be a quarter of the size of an MI sample in the current layer.

The presented patch-based solution is chosen due to its simplicity and effectiveness for tackling the proposed problem. Better solutions might still be formulated, for instance, by adding an edge-preserving regularizer term in (6.2) or by incorporating supplementary data (such as depth, ray-space, and 3D model) into the bitstream. However, this is left for future work.

6.3.2.3 Texture Synthesis

Once the best patch ϕ_e^{best} is obtained by solving (6.2), 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



(a) (b)

Figure 6.10 Examples of two portions of the ILR pictures constructed for each enhancement layer defined in Figure 6.5: (a) the ILR picture that may be used to predict the corresponding original picture depicted in Figure 6.5b; and, (b) the ILR picture that may be used to predict the original picture shown in Figure 6.5c

iterations). Thus, at each iteration, the values of ϕ_e and **B** are updated from the values found in the previous iteration. For the experimental results in Section 6.4, the number of iterations varies from 1 up to 4 (in frames with small MI resolution) and up to 28 (for frames with very large MI resolution).

Figure 6.10 illustrates a portion of the resulting enhanced ILR picture for each enhancement layer defined in Figure 6.5.

6.4 Performance Assessment

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

6.4.1 Experimental Setup

To evaluate the performance of the proposed FOVS-LFC codec, six LF test images with different spatial and MI resolutions are considered so as to achieve representative RD results. These are (see Appendix A): *Fredo*, *Seagull*, *Laura*, *Demichelis Spark (frame1)*, *Robot 3D (frame 54)*, and *Plane and Toy (frame 123)*. The original test images were processed and cropped so as to have all MIs with integer number of pixels, and they were then converted to the Y’CbCr 4:2:0 color format.

To generate the content for each hierarchical layer, *l*, in the scalable codec, a centralized texture sample with size $\left(2^{l+2}\times 2^{l+2}\right)$ is selected from each MI in the LF image data so as to support FOV scalability. These squared texture samples with power of two size were here chosen to better fit into the CTU and PU partition patterns of HEVC [2]. However, these texture samples can be generalized for any sample size and aspect ratio for the proposed scalable codec.

Hence, for each tested LF image, the number of hierarchical layers is given by $\lceil 1/2 \times \log_2 MI_{resol.} \rceil$, where $MI_{resol.}$ is the resolution of each MI in the LF raw image (see this information in Appendix A). Finally, the resolution of each frame in a hierarchical layer l is given by:

- **Fredo, Seagull, and Laura:** $(2^{l+2} \times 96) \times (2^{l+2} \times 72)$ up to 7104×5328 (in the highest enhancement layer).
- **Demichelis Spark:** $(2^{l+2} \times 75) \times (2^{l+2} \times 41)$ up to 2850×1558 (in the highest enhancement layer).
- **Robot 3D, and Plane and Toy:** $(2^{l+2} \times 68) \times (2^{l+2} \times 38)$ up to 1904×1064 (in the highest enhancement layer).

6.4.2 Test and Evaluation Conditions

For the experiments performed in this section, the following test conditions are considered:

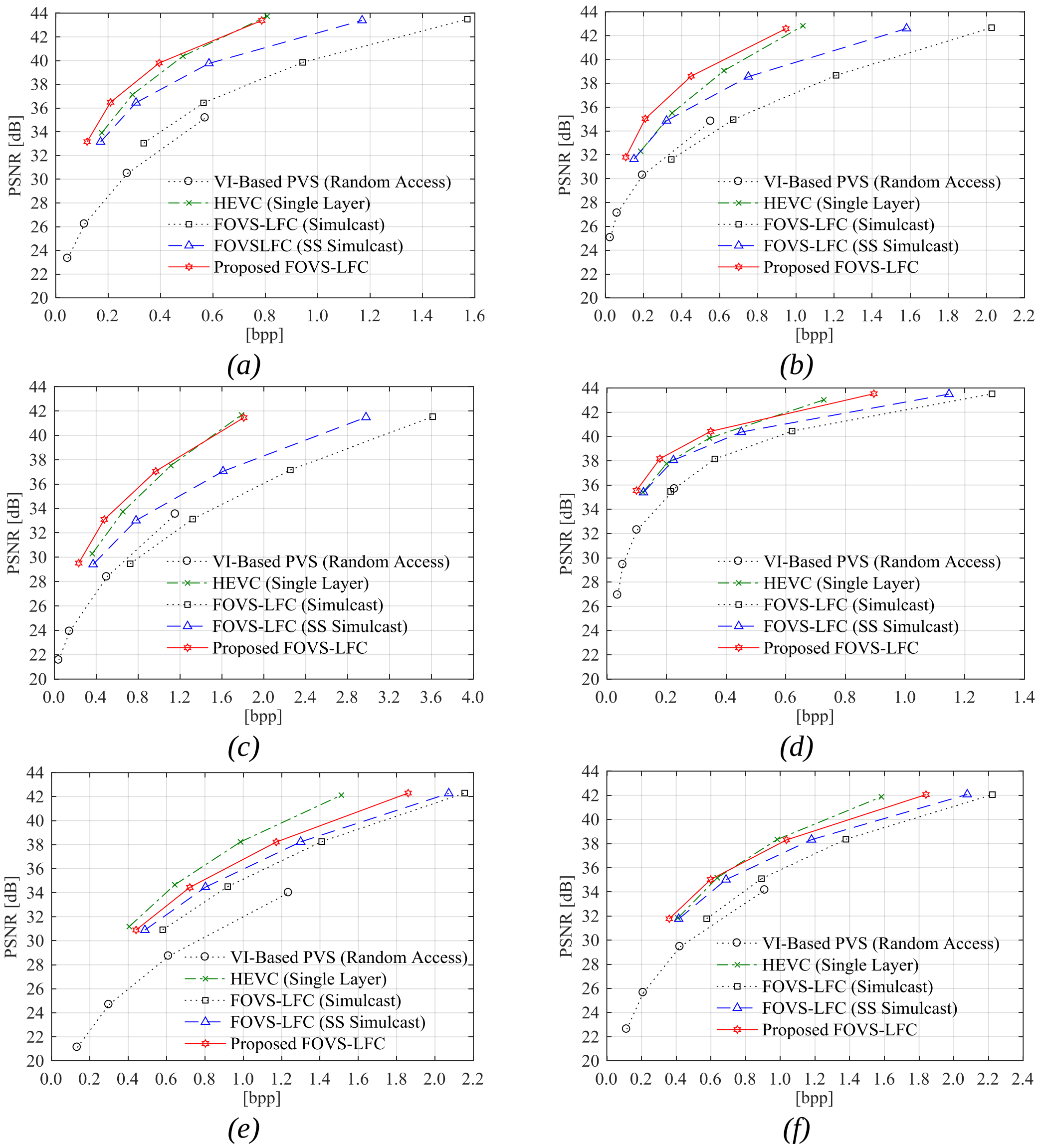
- **Codec Software Implementation** – For these tests, the MV-HEVC reference software version 12.0 [214] is used as a benchmark, as well as the base software for implementing the proposed FOVS-LFC codec.
- **Coding Configuration** – Each of the LF test images is encoded using four different QP values: 22, 27, 32, and 37, according to the common test conditions defined in [176]. The same QP value is used for coding all hierarchical layers.
- **RD Evaluation** – For evaluating the overall RD performance of the proposed FOVS-LFC codec, the distortion is calculated by taking the average luma Mean Squared Error ($\overline{\text{MSE}}$) over the frames from each hierarchical layer, and, then, converting it to the PSNR. The rate is presented in 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 data.
- **Alternative Objective Quality Metrics** – In addition, to analyze the performance in terms of the quality of views synthesized from the reconstructed scalable LF content, the distortion is also measured in terms of the Rendering-dependent PSNR (as in (4.8)) and SSIM (as in (4.9)) metrics. To have a representative number of rendered views, a set of 25 views are rendered from equally distributed viewpoint positions and corresponding to angular information from at least three different hierarchical layers. Moreover, for each of these views, two planes of focus are selected, one to have the main object in focus and another to have the background in focus. In summary, the average PSNR and SSIM values over a total of 50 rendered views are used here as the objective quality metrics. The standard deviation for each result is also presented as a measure of the confidence in the presented average values. For rendering the views, the Basic Rendering algorithm (proposed in [10]) is used.

6.4.3 Benchmark Coding Solutions

The next subsections present and analyze the performance of the proposed FOVS-LFC codec. For this, the following benchmark solutions are compared:

- **FOVS-LFC (Simulcast):** This solution corresponds to the benchmark for the simulcast case, where all frames from each hierarchical layer are coded independently as Intra frames with the standard MV-HEVC solution, using All Intra, Main configuration [176].
- **FOVS-LFC (SS Simulcast):** In this case, each frame 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 intra prediction) but also the non-local spatial correlation between neighboring MIs (with SS prediction). Since, when using the SS prediction, each scalable layer is still coded independently (from each other), 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 (as discussed in Section 6.1.4). Similarly to what has been proposed in [92, 175], a PVS of VIs is coded using HEVC using the Low Delay P [176] configuration. However, in order to fairly compare this solution with the coding in the proposed FOVS-LFC solution, the QP values are kept the same for all VIs in the PVS. Various VI scanning orders were tested (i.e., raster, parallel, zigzag, and spiral), but only the spiral order is presented as it achieved the best RD performance.
- **VI-Based PVS (Random Access)** – In this case, the PVS of VIs scanned in spiral order is encoded using HEVC using Random Access [176] 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.
- **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 [2]. Since the proposed FOVS-LFC codec provides a HEVC-compliant base layer, this solution is used as the benchmark for non-scalable LF coding so as 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.

For the proposed FOVS-LFC, the base layer is encoded as an Intra frame and the remaining enhancement layers are coded as Inter B frames.



(a) *(b)*

(c) *(d)*

(e) *(f)*

Figure 6.11 RD performance for each test image: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark, (e) Robot 3D, and (f) Plane and Toy

6.4.4 Overall FOVS-LFC codec RD Performance

Figure 6.11 shows the RD performance of the proposed FOVS-LFC codec in terms of PSNR versus bpp, and Tables 6.1 and 6.2 present the BD results [210] in terms of PSNR and rate with respect to the benchmarks solutions. From these results, the following conclusions can be derived:

- **Comparison with FOVS-LFC (Simulcast)** – As shown in Figure 6.11 and Table 6.1, the proposed FOVS-LFC RD performance is significantly better than the FOVS-LFC (Simulcast) for all tested images, with average BD gains of 3.84 dB and 46.04 % of bit savings. The gains are much more expressive for test images with higher MI resolution,

Table 6.1 RD performance (using BD metrics [210]*) of the proposed FOVS-LFC codec regarding the simulcast solutions (QP values 22, 27, 32, and 37)*

LF Image	**FOVS-LFC (Simulcast)**		**FOVS-LFC (SS Simulcast)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**
Fredo	5.32	-59.70	2.07	-32.27
Seagull	5.69	-64.63	2.26	-38.40
Laura	**5.95**	**-59.62**	2.87	-39.59
Demichelis Spark (frame 1)	2.26	-45.92	0.94	-23.22
Robot 3D (frame 54)	1.69	-18.55	0.82	-9.83
Plane and Toy(frame 123)	2.15	-27.85	0.83	-12.34
Average	**3.84**	**-46.04**	**1.63**	**-25.94**

where the BD-PSNR gain goes up to 5.95 dB (for *Laura*) with 59.62 % of bit savings. These gains are justified by the efficiency in exploiting the redundancy between the layers (using the proposed IL coding tools), as well as the efficiency in exploiting the correlations within a single enhancement layer (using the SS prediction).

- **Comparison with FOVS-LFC (SS Simulcast)** – Comparing this solution with the complete FOVS-LFC in Figure 6.11 and Table 6.1, it can be seen that the proposed FOVS-LFC has better RD performance with average BD gains of 1.63 dB and 25.94 % of bit savings. As expected, it is shown that improved RD performance can be attained by taking advantage of the redundancy in the inter-layer domain.

- **Comparison with PVS-Based Approaches** – It can be seen in Table 6.2 and Figure 6.11 that the proposed FOVS-LFC solution architecture presents significantly better RD performance than the PVS-based arrangement of the VIs, for both tested configurations (Low Delay P and Random Access). The average BD gains of the proposed FOVS-LFC are 5.72 dB and 64.23 % of bit savings when compared to the VI-based PVS (Low Delay P), and 4.67 dB and 56.59 % of bit savings when compared to VI based PVS (Random Access).

- **Comparison with HEVC (Single Layer)** – As shown in Table 6.2, the proposed FOVS-LFC solution presents a slightly better performance than the non-scalable HEVC (Single Layer) with average BD gains of 0.39 dB and 7.76 % of bit savings. This shows that by using efficient prediction schemes between the hierarchical layers, it is possible to achieve scalable LF coding at no additional cost (in average) in comparison to the non-scalable HEVC (Single Layer). Notice that the larger the resolution of the tested image is, the better is the performance of the scalable proposed FOVS-LFC solution (up to 1.57 dB with 25.73 % of bit savings). For the tested images with smaller resolution (*Plane and Toy* and *Robot 3D*), the non-scalable HEVC (Single Layer) solution presented the best performance. However, as it will be further

Table 6.2 RD performance (using BD metrics [210]*) of the proposed FOVS-LFC codec regarding PVS-based and HEVC (Single Layer) solutions (QP values 22, 27, 32, and 37)*

LF Image	**VI-based PVS (Low Delay P)**		**VI-based PVS (Random Access)**		**HEVC (Single Layer)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
Fredo	7.43	-74.42	7.16	-71.49	0.79	-13.86
Seagull	4.80	-68.07	4.47	-65.67	**1.57**	**-25.73**
Laura	5.56	-62.85	4.81	-57.60	0.73	-11.47
Demichelis Spark (frame 1)	4.96	-68.12	3.57	-55.25	0.56	-13.98
Robot 3D (frame 54)	6.17	-54.44	4.56	-46.48	-1.27	17.35
Plane and Toy(frame 123)	5.43	-57.45	3.44	-43.06	-0.08	1.12
Average	**5.72**	**-64.23**	**4.67**	**-56.59**	**0.39**	**-7.76**

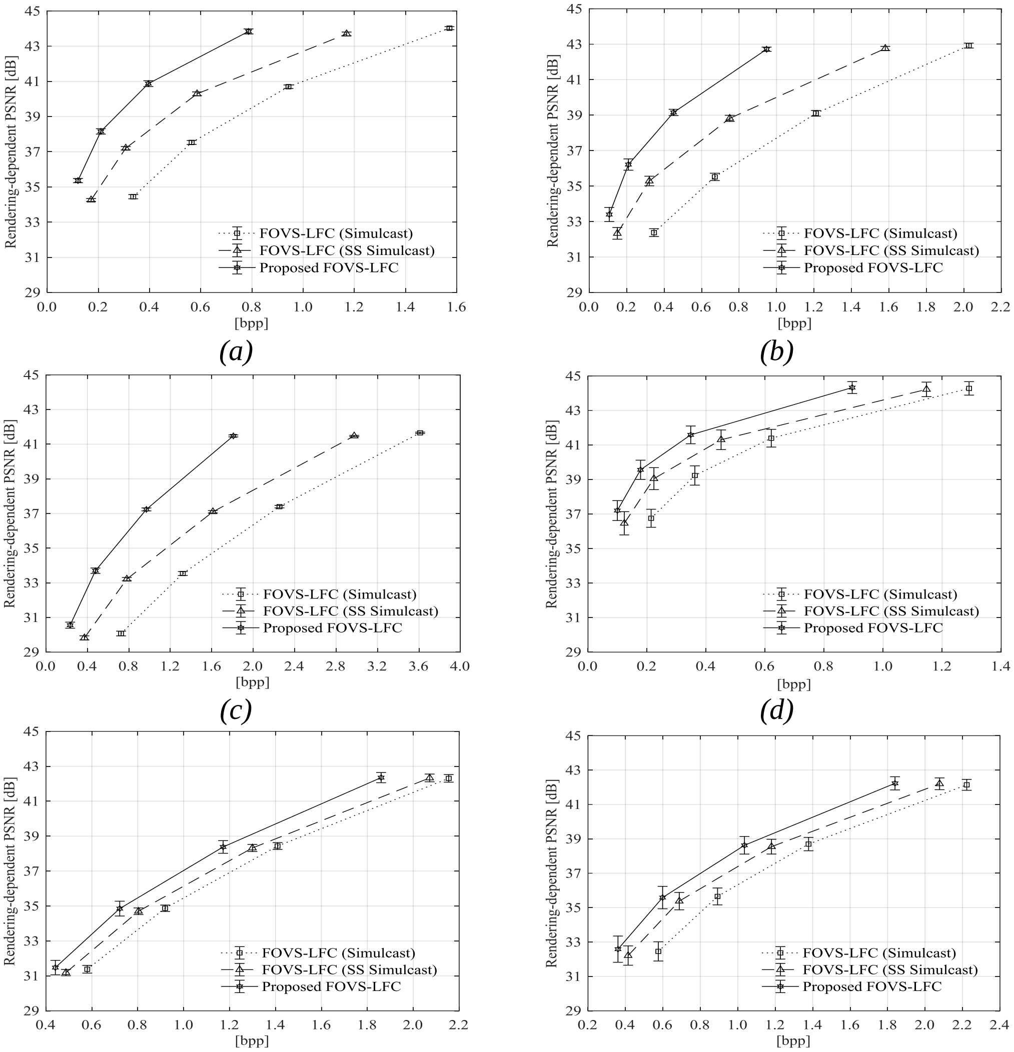
discussed in the next subsections, a major advantage of the proposed FOVS-LFC solution is the ability to provide a bitstream that is flexible in terms of the LF interaction functionalities and that enables rendering high quality views in all hierarchical layers at no additional bitrate cost (in average).

6.4.5 Quality of Rendered Views

In order to assess the performance of the proposed scalable coding architecture regarding the quality of rendered views, the RD performance of the proposed FOVS-LFC is here presented in terms of the Rendering-dependent PSNR (see Figure 6.12) and SSIM (see Figure 6.13) metrics over a set of rendered views (as explained in Section 6.4.1) and compared to the simulcast case. This comparison aims at analyzing the RD improvements due to the inclusion of the proposed IL prediction schemes into the proposed LF enhancement layer codec (illustrated in Figure 6.7). For this reason, other coding architectures (i.e., the PVS-based solutions and the non-scalable HEVC (Single Layer)) are not presented in this analysis. However, a supplementary comparison with respect to HEVC (Single Layer) will be presented in Section 6.4.6.

From these results, the following conclusions can be drawn:

- **Overall Performance** – Considering the RD results for all metrics used (as shown in Figures 6.11-6.13), the proposed FOVS-LFC outperforms both simulcast solutions – i.e., FOVS-LFC (Simulcast) and FOVS-LFC (SS Simulcast) – with significant gains. 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 6.1.4), there is no discrepancy between the quality of the entire LF image (see Table 6.1 and Figure 6.11) and the quality of rendered views (see Figures 6.12-6.13) when using the proposed scalable solution with FOV scalability.



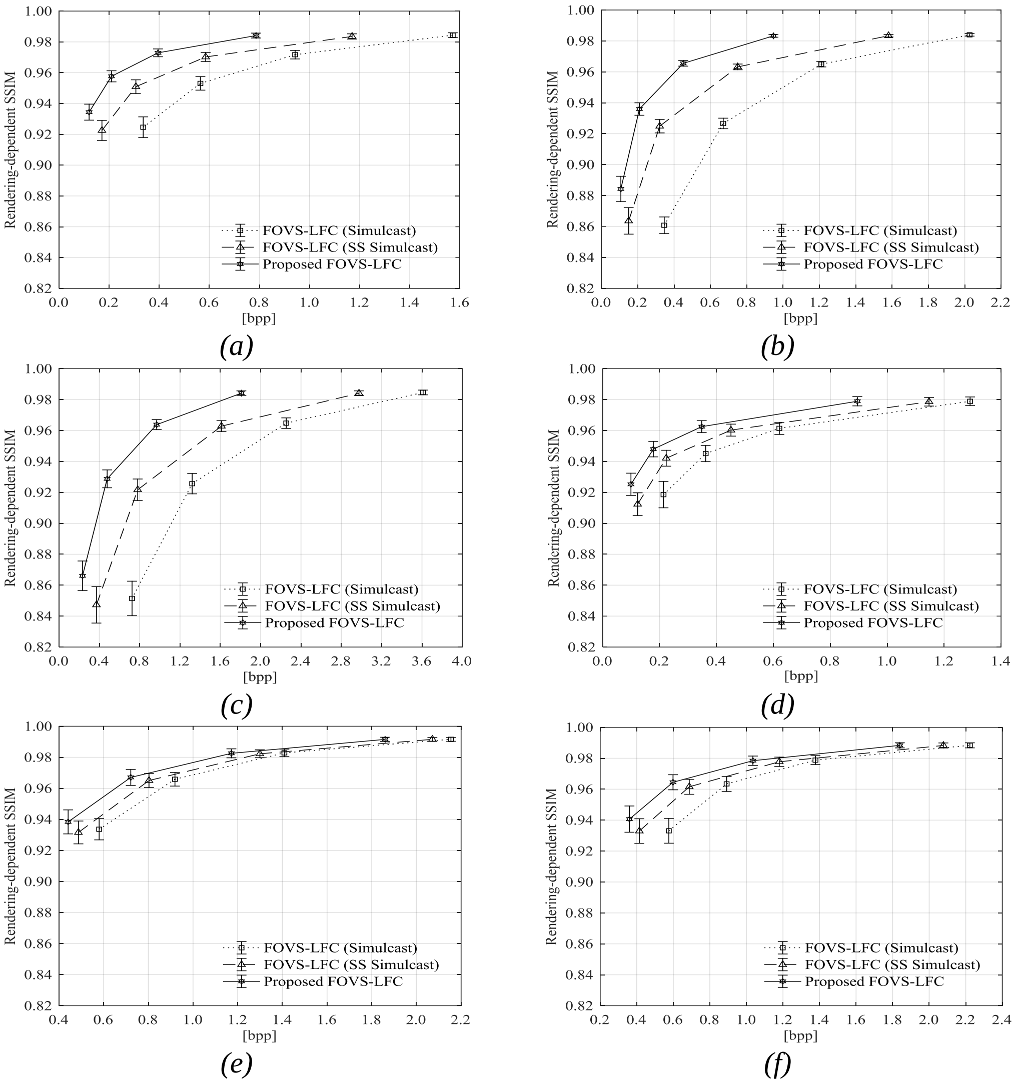
(a) *(b)*

(c) *(d)*

(e) *(f)*

Figure 6.12 RD performance (in terms of PSNR versus bpp) for a set of rendered views from image: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark, (e) Robot 3D, and (f) Plane and Toy

- **Comparison between Different Objective Quality Metrics** – Comparing again the results in Figures 6.11-6.13, it can be seen that there is a similar trend in the relative coding performance using the three different quality metrics. Regarding the standard deviation values presented in Figures 6.12 and 6.13, it can be observed that LF images with smaller MI resolution (i.e., *Demichelis Spark*, *Robot 3D* and *Plane and Toy*) present higher variation in PSNR values for all coding solutions. A more careful analysis of the PSNR results for each rendered view showed that a set of views rendered from viewpoint positions near to the border of the MI present larger variation in PSNR values than a set of views from middle viewpoint positions. This may be due to the vignetting that appears in the border of each MI. As an illustrative example, considering the PSNR-based results for *Plane and Toy* using the proposed FOVS-LFC solution with QP 37, the standard deviation value goes from 0.18 dB over the middle

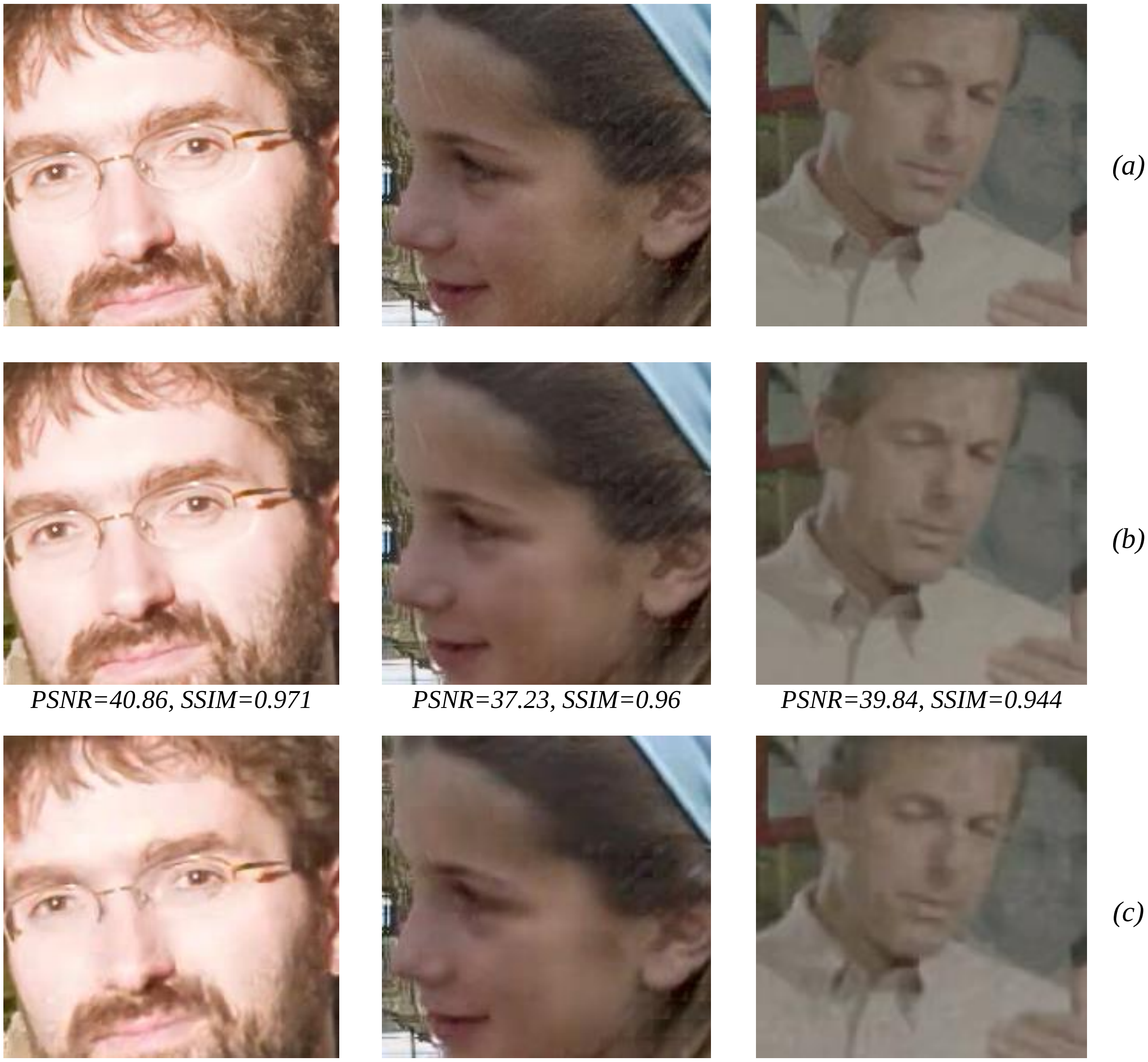


(a) *(b)* *(c)* *(d)* *(e)* *(f)*

Figure 6.13 RD performance (in terms of SSIM versus bpp) for a set of rendered views from image: (a) Fredo, (b) Seagull, (c) Laura, (d) Demichelis Spark, (e) Robot 3D, and (f) Plane and Toy

views to 0.59 dB over the border views. The same happens for the SSIM values for the tested LF images.

- **Comparison by Visual Quality Inspection of Rendered Views** – Figure 6.14 illustrates a portion of a central view from three different tested images coded with the proposed FOVS-LFC compared to the FOVS-LFC (Simulcast) (using the same bpp) to illustrate how much it is possible to improve the visual quality with respect to the simulcast case. From these results, it can be seen that the proposed FOVS-LFC solution also improves the visual quality of the rendered views. Similar conclusions were observed for the other LF test images. As illustrated in this figure, coding artifacts are significantly more visible in the FOVS-LFC (Simulcast) case, such as, for example, in the glasses and hair of *Fredo*, in the eyes of *Laura*, and all over the face of *Demichelis Spark*. Moreover, there is a similar trend between the compared quality metrics and the visual quality of rendered views. In addition to this, it can be observed that the SSIM-based metric generally presented very high SSIM values for all coding solutions (for the bpp ranges used) and small variations in the SSIM values compared to the variations in the perceived quality (e.g., for the views from *Demichelis Spark* in Figure 6.14 (b and c) the difference in SSIM values can be only seen in the second decimal place). This may indicate that SSIM is not as sensitive as the PSNR to the perceived variances in quality for this application scenario.



PSNR=40.86, SSIM=0.971 *PSNR=37.23, SSIM=0.96* *PSNR=39.84, SSIM=0.944*

PSNR=35.14, SSIM=0.944 *PSNR=32.12, SSIM=0.894* *PSNR=42.09, SSIM=0.960*

Figure 6.14 Comparison of a portion from the central 2D view of Fredo (left), Laura (middle), and Demichelis Spark (right) when rendering from: (a) the original image; (b) encoded and reconstructed frame using proposed FOVS-LFC; and (c) encoded and reconstructed frame using FOVS-LFC (Simulcast). The corresponding SSIM and PSNR values (Rendering-dependent) are also shown for the three test images at around at 0.4 bpp for Fredo, 1 bpp for Laura, and 0.4 bpp for Demichelis Spark

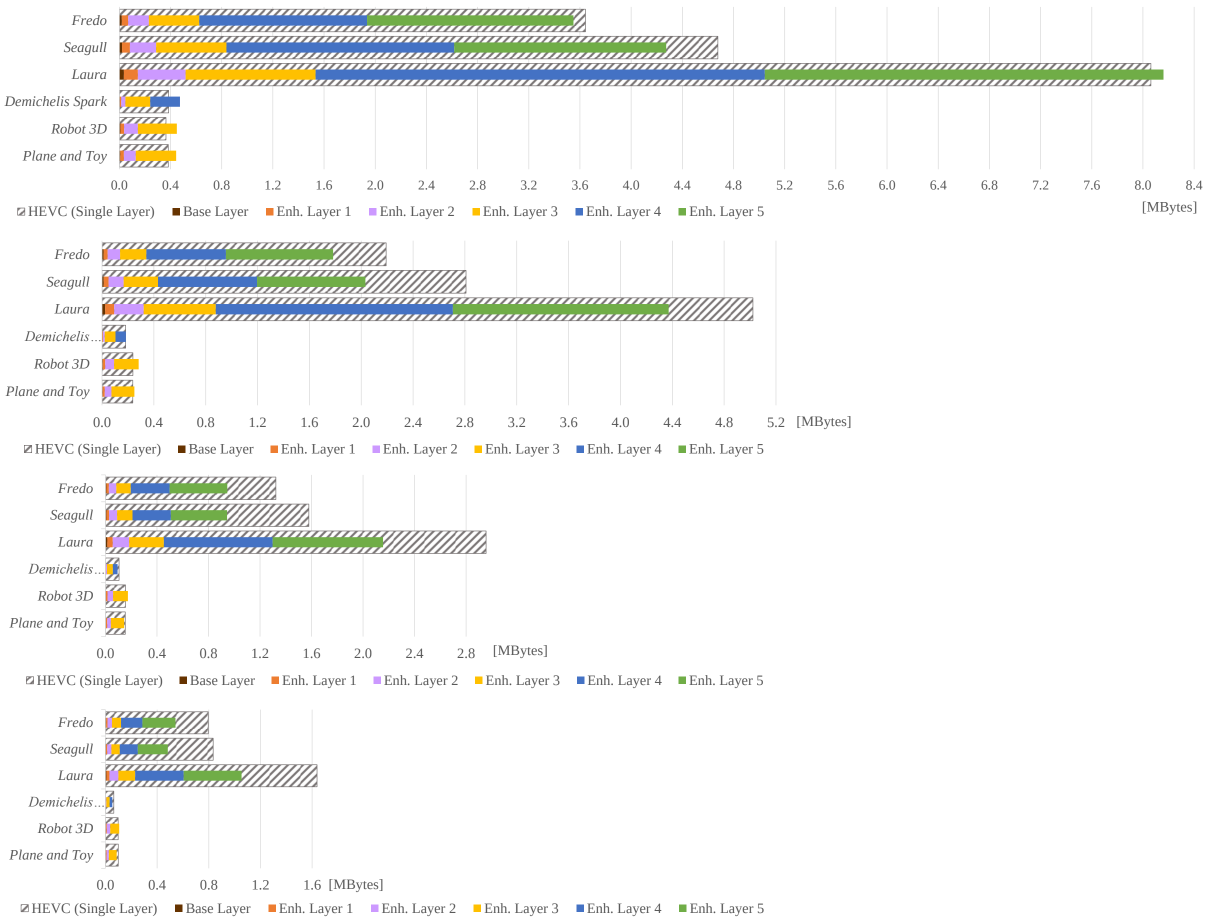
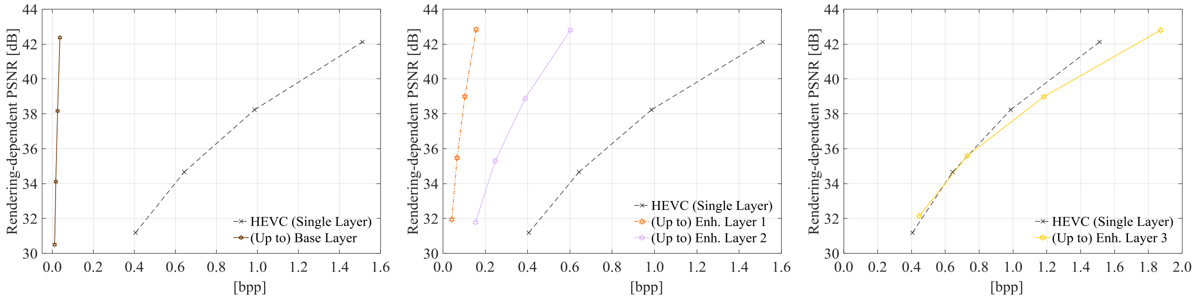


Figure 6.15 Coding bits (in Mbytes) for each scalable layer using the FOVS-LFC codec, and comparing against the non-scalable benchmark solution HEVC (Single Layer) for the QP values (top to bottom): 22, 27, 32, and 37

6.4.6 Performance for Different Application Scenarios

In order to further discuss the usability of the proposed scalable coding solution, the performance of the proposed FOVS-LFC 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 (see Figure 6.1). For each of the considered scenarios, the corresponding RD performance of proposed FOVS-LFC is compared to the benchmark HEVC (Single Layer), in which scalability is not supported, so as to analyze the advantages of the proposed FOVS-LFC solution in terms of the flexibility enabled in the bitstream.

To facilitate this analysis, Figure 6.15 firstly illustrates the needed bits for encoding each of the scalable layers using the proposed FOVS-LFC compared to the bits needed for encoding the entire LF raw data using the non-scalable solution HEVC (Single Layer), for all 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.



(a) *(b)* *(c)*

Figure 6.16 RD performance 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. In all scenarios, rate is given by the total number of bits need for encoding up to the corresponding hierarchical layer divided by the number of pixels in the entire (raw) LF data. Moreover, distortion is given in terms of Rendering-dependent PSNR for a central view rendered from the coded LF data in the corresponding hierarchical layer

Nevertheless, this analysis will consider the worst case scenario where the proposed FOVS-LFC solution does exceed the encoding bits of 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, Figure 6.16 shows the RD performance for the test image *Robot 3D* (frame 54), in terms of PSNR of a central rendered view and bpp for each of the following scenarios:

- **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 (see Figure 6.1), 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 Figure 6.15, the base layer corresponds to a very small percentage of the complete scalable bitstream. Therefore, the RD efficiency of the proposed FOVS-LFC solution would greatly increase, as shown in Figure 6.16a.

- **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 (see Figure 6.1). 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 Figure 6.16b. In both

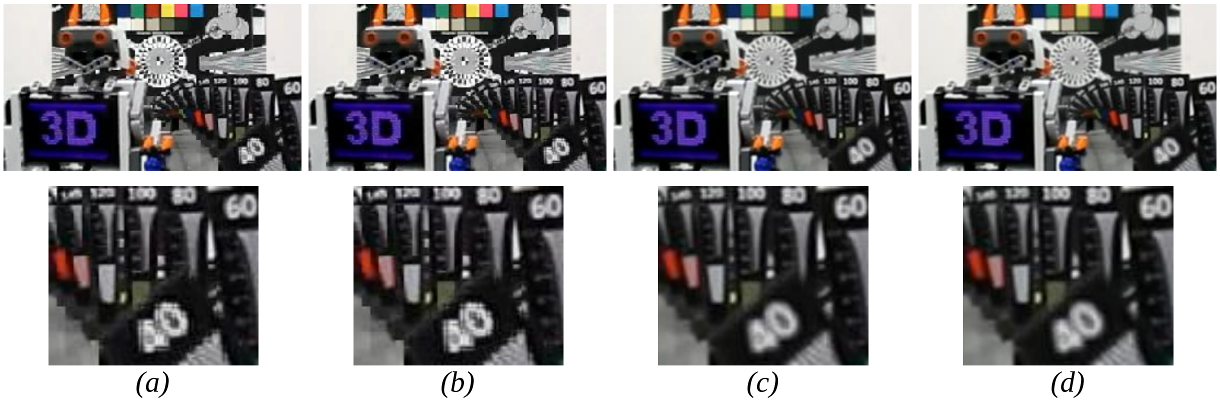


Figure 6.17 Example of views for rendered from Robot 3D (frame 54) when using the proposed FOVS-LFC codec (with QP 22). Each image corresponds to a different hierarchical layer: (a) base layer; (b) enhancement layer 1; (c) enhancement layer 2; and (d) enhancement layer 3. It is possible to observe how the larger angular information in higher layer allows having richer refocusing effects when manipulating the rendered views. This can be noticeable mainly by the blur at the out of focus areas (bottom) of the view

cases, it is still possible to significantly improve the coding efficiency compared to the HEVC (Single Layer). Figure 6.17 illustrates a portion of views rendered from reconstructed frames in each scalable layer for the tested image *Robot 3D* (frame 54). All views were rendered from the same viewpoint position. As expected, the richer angular information in higher layers (from Figure 6.17a to Figure 6.17d) allows the user to have larger degrees of freedom in manipulation (e.g., refocusing and changing the perspective). However, comparing Figure 6.17b and Figure 6.17c with Figure 6.17d, it can be seen that in Figure 6.17c the user may not need to decode the complete bitstream to have rendered views with similar perceived results to Figure 6.17d.

- **Scenario 3 (Full Interaction Capabilities)** – This scenario supports LF applications in which the user demands full rendering capabilities and visualization with maximum angular information (see Figure 6.1). 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. Figure 6.16c 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 6.2 shows that for other images the proposed FOVS-LFC outperforms HEVC (Single Layer) with bit savings of up to 25.73 % for LF images with larger resolutions, and notice that, 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.

6.4.7 Computational Complexity

As usually observed, the significantly better flexibility of the FOVS-LFC codec comes with the price of additional computational load when compared to HEVC intra prediction.

Table 6.3 FOVS-LFC Encoding and Decoding Execution Time

Encoding Time [s]							
LF Image	**Proposed FOVS-LFC Layers:**						**HEVC (Single Layer)**
	Base	**1**	**2**	**3**	**4**	**5**	
Plane and Toy	0.2	53.3	242.6	862.0	-	-	127.5
Seagull	0.7	200.7	779.1	3020.6	11583.2	19885.1	389.7
Decoding Time [s]							
LF Image	**Proposed FOVS-LFC Layers:**						**HEVC (Single Layer)**
	Base	**1**	**2**	**3**	**4**	**5**	
Plane and Toy	0.0	0.4	4.2	12.1	-	-	0.3
Seagull	0.0	1.2	3.1	27.9	182.9	427.6	4.7

Regarding the SS and the IL compensated predictions, the encoder and decoder computational complexity is conceptually the same as for HEVC inter prediction. Detailed information about HEVC computational complexity can be found in [216]. Concerning the exemplar-based direct IL prediction, the encoder complexity is similar to HEVC inter prediction, but the decoder complexity is increased since coding blocks that use this type of prediction will have to estimate the direct IL vectors. For these coding blocks and considering a search window $\mathbf{W} \in \mathbb{R}^{N \times N}$, the decoder will need to perform up to N^2 additional SAD computations (in a p-pixels block as seen in (6.1) and Figure 6.8). Regarding the patch-based ILR picture construction, encoder and decoder complexity are similar, and the algorithm is employed once for each hierarchical layer. For instance, suppose that the picture is formed of an $M \times L$ grid of MIs, and, then, it can be divided into $4 \times M \times L$ patches. It was seen in Section 6.3.2 that, for each patch, the minimization in (6.2) is performed in a neighborhood of K MIs. Therefore, in each algorithm iteration, up to $4 \times M \times L \times K$ SAD computations are performed (in an n-pixels block as seen in (6.2) and Figure 6.9). Notice that the second term in (6.2) can be computed while re-organizing the texture from the reference layer.

To complete this analysis, Table 6.3 gives the execution time for encoding and decoding each FOVS-LFC hierarchical layer (using a machine with an Intel Core i7-4700HQ processor at 2.40 GHz) compared to HEVC (Single Layer), which uses only Intra prediction. These values are shown for two LF images with the smallest (*Plane and Toy*) and the largest (*Seagull*) LF raw and MI-grid resolutions. It can be seen that the computational complexity is considerably smaller in the lower FOVS-LFC layers. Moreover, the largest execution time for both images (Layer 3 for *Plane and Toy*, and Layer 5 for *Seagull*) corresponds to the worst case observed in terms of the number of iterations for the patch-based ILR picture construction.

It is worth noting that scaling the complexity load is also advantageous since the user may not need to decode and process the complete bitstream.

6.5 Final Remarks

This chapter proposed a flexible and efficient scalable coding framework for emerging LF imaging applications that provides a novel type of scalability, here referred to as FOV scalability. The proposed FOVS-LFC codec comprises an HEVC backward compatible base layer and a flexible number of LF enhancement layers, which are coded using the SS prediction combined with two novel IL prediction schemes for improving the compression performance.

The proposed scalable coding architecture satisfies many of the current requirements for emerging image and video technologies, being easily adaptable to various use case scenarios demanding richer and immersive visualization. Moreover, experimental results showed that the proposed FOVS-LFC solution was able to achieve significantly better rate-distortion performance compared to the benchmark scalable solutions. Furthermore, the proposed scalable design provided flexibility in the rendering functionalities that emerge from LF imaging applications with bit savings comparable to the non-scalable benchmark HEVC. In addition to this, it was shown that the compressed rendered views presented high quality in all hierarchical layers.

Furthermore, future work was briefly discussed and includes the investigation of opportunities to enhance the proposed exemplar-based IL prediction, by adding a regularizer term in the proposed problem formulation, as well as by incorporating supplementary data (such as depth, ray-space, and 3D model) into the scalable bitstream.

Chapter 7

Achievements and Future Directions

To conclude this Thesis, this chapter summarizes the main achievements that were reached during its course and it also identifies some possible future research directions.

7.1 Summary of the Thesis Achievements

Microlens-based LF imaging (also known as integral, holoscopic, and plenoptic imaging) has recently risen up as feasible and promising technology for providing richer visual experiences with new degrees of freedom in terms of content production and manipulation. Moreover, there has been an increasing number of camera and display manufacturers seeking to improve the performance of this technology, with some products already having reached the consumer market.

Acknowledging the potential of this technology, JPEG and MPEG standardization bodies have recently decided to initiate efforts to support, in a near future, LF imaging applications in a standardized way. Among the requirements that have been discussed, deploying new coding solutions for LF content in order to efficiently deliver it to end-users is one of utmost importance. Recognizing this, this Thesis has addressed this specific concern by developing and evaluating various solutions for efficient LF content coding.

In this context, the first achievement of this Thesis was the development of an efficient LF coding solution based on HEVC and using SS compensated prediction. This solution was presented in Chapter 4. The main advantage of the proposed solution was to support high RD coding performance while being agnostic to the optical system used for the LF acquisition (e.g., microlens size, focal length and distance of the microlenses to the image sensor). In the proposed solution, the (raw) LF image could be directly encoded without needing any previous calibration and/or conversion to a different format. For achieving high RD efficiency, a Bi-SS estimation was proposed, in which the current block (being coded) could be predicted from two candidate blocks that were jointly estimated from the same search window (inside the SS reference) by using a locally optimal rate-constrained algorithm [58]. In addition, a novel SS vector prediction, named MIVP, was also used to achieve further bit savings. It was shown that the SS compensated prediction was able to significantly improve the RD performance when compared to HEVC Still Picture Profile, while keeping the encoding/decoding complexity and memory load comparable to HEVC inter coding. Moreover, jointly estimating the two candidate blocks for Bi-SS prediction led to further RD improvements when compared to the case where only one candidate block is estimated (referred to as Uni-SS), as well as compared to the case where the two candidate blocks are independently estimated (referred to as Restricted-SS). Moreover, it was shown that (see Appendix B) significantly better RD performance compared to HEVC Still Picture Profile was also possible for LF images captured with the Lytro Illum traditional LF camera.

The second Thesis achievement was related to the need for having a scalable LF coding solution that provides backward compatibility to legacy display devices. This is an important requirement for enabling faster deployment of new LF imaging applications in the consumer market, which was tackled in Chapter 5. In the proposed solution, a three-layer scalable coding architecture was used, in which the two lower layers supported coding of 2D, stereo, or multiview versions of the same LF content, while the highest layer supported the coding of the entire LF content. For improving the RD coding performance in the highest layer, an IL prediction was proposed and combined with the SS compensated prediction. It was shown that the proposed scalable LF coding solution enabled backward display compatibility while still achieving, in most of the cases, better RD performance than HEVC Still Picture Profile. It was also seen that the proposed combined prediction achieved significant RD performance improvements compared to the simulcast cases.

Finally, the third achievement of this Thesis concerns the requirement to support richer and flexible interactive functionalities in LF imaging applications. This requirement was tackled in Chapter 6. To accomplish this, an LF coding solution with FOV scalability was proposed, which was built upon a multi-layer scalable coding architecture. The FOV scalability supported progressively richer forms of the same content by hierarchically organizing the angular information of the captured LF content. The proposed FOV scalable coding architecture provided backward compatibility to HEVC in the base layer, and defined one or more LF enhancement layers containing progressively richer angular information, which were them encoded with the proposed LF enhancement layer encoder. To improve the RD coding efficiency in an LF enhancement layer, two novel IL prediction schemes were proposed, which were also combined with the SS compensated prediction. It was shown that the proposed scalable architecture was able to provide flexibility in the coding and rendering functionalities that emerge from LF imaging applications, while achieving RD performance comparable to the non-scalable coding with HEVC Still Picture Profile. Furthermore, the proposed scalable coding solution provided high quality rendered views in all hierarchical layers.

7.2 Future Research Directions

This is a very exciting time for working with LF imaging technologies, and there are still many future research challenges that need to be solved for enabling successful LF imaging applications and services. These challenges are related to all stages of the LF processing chain, comprising the development of new cameras and displays, novel representation models associated to efficient coding solutions, as well as powerful rendering algorithms and quality assessment methodologies for LF imaging experiences.

This Thesis has been focused on the development of new and efficient coding solutions. Therefore, this section discusses some work items that naturally follow from the topics discussed in this Thesis, and which still deserve to be pursued in the future.

- **Design of Optimized Filters for SS Compensated Prediction** – It was seen, in Chapter 4, that the proposed Bi-SS prediction was able to improve the RD performance (with respect to the Uni-SS prediction and the Restricted-SS prediction) by simply averaging two jointly estimated predictor blocks. However, it is still possible to achieve further RD performance by using an optimized filter for bi-prediction. In fact, the theoretical analysis in [211] indeed suggests that the optimized filter becomes increasingly important for higher residual noise levels, and then the optimal filter is able to improve the RD performance by reducing the noise in the compensated prediction signal [211]. Therefore, proper algorithms for estimating an optimal set of weighting coefficients for Bi-SS prediction are still needed, and deserve future considerations.
- **Development of Depth/Disparity Estimation Algorithms from LF Images** – Estimating depth/disparity in real-world LF images (captured with a microlens-based LF camera) is a very challenging problem that still needs further research in the future. Apart from the challenges commonly faced by any depth estimation algorithm – such as dealing with the presence of occlusion and transparent, reflective, and specular surfaces – LF cameras typically present a narrow baseline and small disparity ranges (e.g., the disparity range for Lytro first generation LF cameras is reported to be between -1 to 1 pixels in [89]). Consequently, accuracy in the depth/disparity estimation becomes even more challenging [88]. Regarding depth/disparity-assisted coding solutions, the open question that still needs to be further investigated and answered is if the currently available depth/disparity estimation algorithms for LF cameras (e.g., in [86–90]) can be potentially used for diminishing the amount of LF texture that is encoded and transmitted in the LF processing chain. To the best of the author's knowledge, the depth/disparity-assisted coding schemes proposed in the literature (e.g., in [206, 208]) still make use of a not very accurate depth/disparity estimation process, in which a single disparity value is extracted for an entire MI. As a result, the quality of the reconstructed/synthesized texture and, consequently, the

quality of rendered views (i.e., rendered from this reconstructed LF texture data) is still severely affected by these inaccuracies in the estimated disparity/depth.

- **Development of Improved IL Prediction Schemes for Scalable LF Coding** – There are still opportunities to improve the IL prediction schemes that were proposed in this Thesis – namely, the MI refilling-based IL reference construction (proposed in Chapter 5), and the patch-based IL reference construction (proposed in Chapter 6). In both schemes, the problem is to synthesize texture for constructing an efficient IL reference picture. Therefore, more advanced inpainting algorithms proposed in the literature (e.g., in [217]) can still be investigated and adapted to this specific type of content. Another option to be also studied in the future is to make use of the estimated depth/disparity information to improve this inpainting process.

To finalize, it is important to say that, although standardized LF coding solutions are still in an early stage of development, it is likely that, in a near future, the research on this topic will advance significantly – mainly due to the current JPEG Pleno and MPEG Light Field Compression standardization activities. In addition, both JPEG Pleno and MPEG Light Field Compression groups have been issuing calls for LF materials contributions [35, 49]. Thus, it likely that more representative LF datasets will be created in order to extend, for instance, the recently made available Lytro datasets (e.g. in [218, 219]). In addition, it is expected that these datasets may consider not only LF content with different texture characteristics (such as natural, transparent, reflective, specular [218, 219]) but also captured using different LF camera setups (e.g., conventional and focused cameras, different MLA sizes, shapes and packing schemes, and different main lens focal length and aperture sizes). This larger set of parameters will allow having LF content with different characteristics, which may be more representative of particular LF application scenarios. Therefore, it is expected that the LF coding techniques proposed in this Thesis, as well as the ones available in the literature, will be further validated in order to draw more generic conclusions with respect to the RD performance for a larger variety of LF application scenarios.

Appendix A
Light Field Test Content

This appendix illustrates the LF test content that is used throughout this Thesis. In addition, a brief description of the LF content characteristics is provided so as to avoid repeating it in each experimental test.

The content is here separated into LF images, presented in Section A.1, and LF sequences, presented in Section A.2.

It is important to mention that, for the experimental results presented throughout this Thesis, these LF images and video sequences were pre-processed before being coded. For this, the following steps were adopted:

1) The original (raw) LF image in the RGB 4:4:4 color format is converted to Y’CbCr 4:4:4.
2) The Y’CbCr 4:4:4 LF image is subsampled to the Y’CbCr 4:2:0 color format.
3) Finally, the Y’CbCr 4:2:0 LF image is rectified using DCT-based quarter-pixel interpolation filters [220] from HEVC in order to approximate the non-integer MI size to an integer value. In this process, incomplete MIs at the border of the LF image are also discarded.

More recently, the Matlab LF Toolbox [107] has been made publicly available for pre-processing and calibrating LF images captured using Lytro LF cameras [12]. However, there is still no consensus about which algorithms should be adopted.

Moreover, at the time the research work presented in this Thesis was done, there was no common test conditions specified for evaluating LF coding solutions, and representative LF datasets were still being gathered for standardization purposes [35, 49].

A.1 LF Test Images

This section illustrates the natural LF test images (available in [221]), whose main characteristics are presented in Table A.1.

Table A.1 Main characteristics of the LF images in Figures A.1-A.5

LF Image	*Fredo, Jeff, Laura, Seagull, and Zhengyun1*
LF Resolution	7240×5432 (original) and 7104×5328 (rectified)
Camera Setup	Focused (Plenoptic Camera 2.0)
MLA	Rectangular-arranged, squared-based lenses
Microlens	0.5 mm (pitch), 1.5 mm (focal length)
MI Resolution	~74×74 (integer approximation)

A.1.1 Fredo

Fredo is shown in Figure A.2, and its main characteristics are summarized in Table A.1.

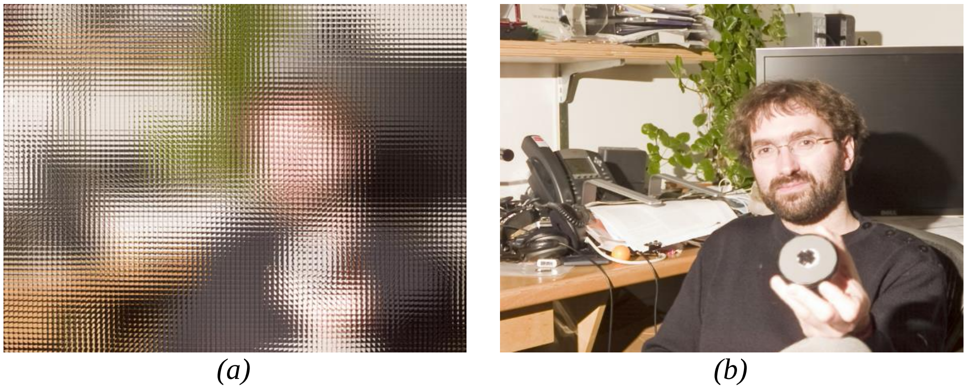

(a) *(b)*

Figure A.1 Fredo LF image: (a) LF image; and (b) Central view rendered from the LF image

A.1.2 Jeff

Jeff is shown in Figure A.2, and its main characteristics are summarized in Table A.1.

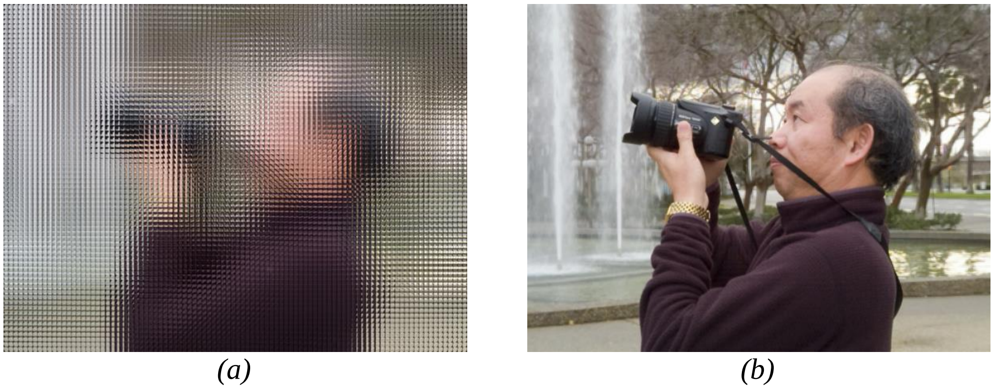
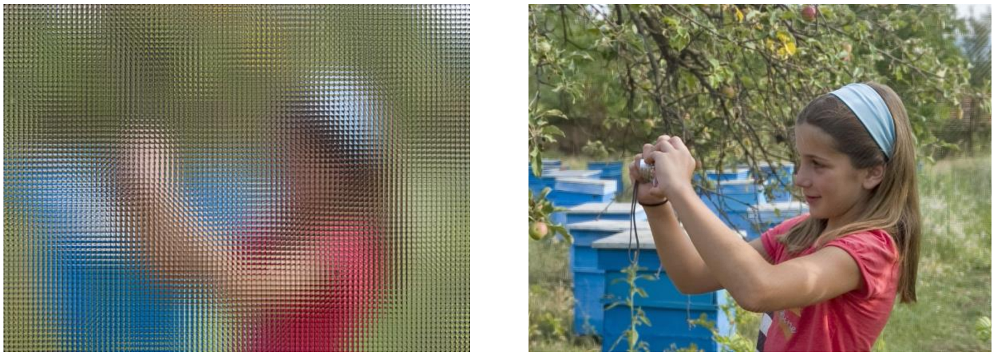

(a) *(b)*

Figure A.2 Jeff LF image: (a) LF image; and (b) Central view rendered from the LF image

A.1.3 Laura

Laura is shown in Figure A.3, and its main characteristics are summarized in Table A.1.

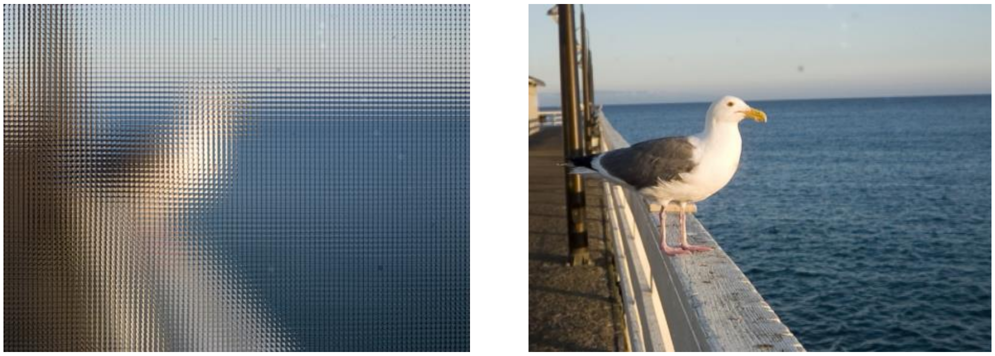


(a) *(b)*

Figure A.3 Laura LF image: (a) LF image; and (b) Central view rendered from the LF image

A.1.4 Seagull

Seagull is shown in Figure A.4, and its main characteristics are summarized in Table A.1.

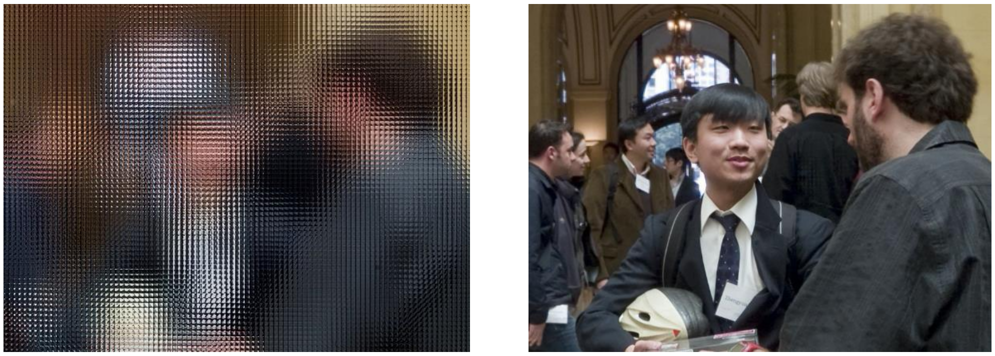


(a) *(b)*

Figure A.4 Seagull LF image: (a) LF image; and (b) Central view rendered from the LF image

A.1.5 Zhengyun1

Zhengyun1 is shown in Figure A.5, and its main characteristics are summarized in Table A.1.



(a) *(b)*

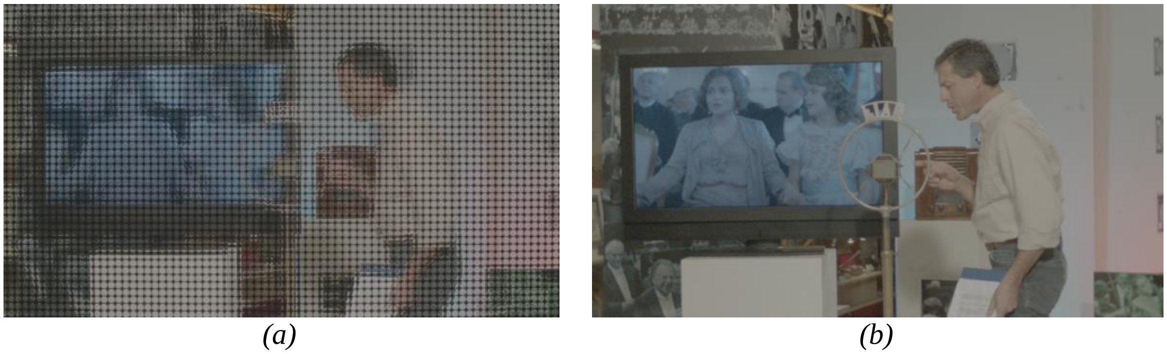
Figure A.5 Zhengyun1 LF image: (a) LF image; and (b) Central view rendered from the LF image

A.2 LF Sequences

This section illustrates the LF test sequences (available in [55]) and their characteristics.

A.2.1 Demichelis Cut

Demichelis Cut is shown in Figure A.6, and Table A.2 summarizes its main characteristics.



(a) *(b)*

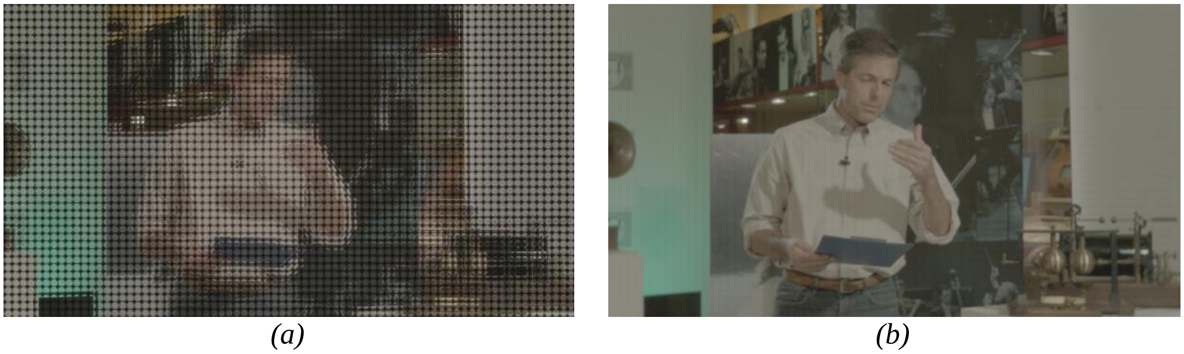
Figure A.6 Demichelis Cut LF sequence (1st frame): (a) LF content; and (b) Rendered Central View

Table A.2 Main characteristics of LF sequence Demichelis Cut

LF Sequence	*Demichelis Cut*
LF Resolution	2880×1620 (original) and 2850×1558 (rectified)
Frame Rate	25 Hz (150 frames)
Camera Setup	Focused (Plenoptic Camera 2.0)
MLA	Rectangular-arranged, semi-spherical lenses
Microlens	0.3 mm (pitch), 2.2 mm (focal length)
MI Resolution	38×38
Description	Natural studio scene with a presenter reading the script. Behind him, in addition to the static background, there is a large monitor showing a video related to what is being read.

A.2.2 Demichelis Spark

Demichelis Spark is shown in Figure A.7, and Table A.3 summarizes its main characteristics.



(a) *(b)*

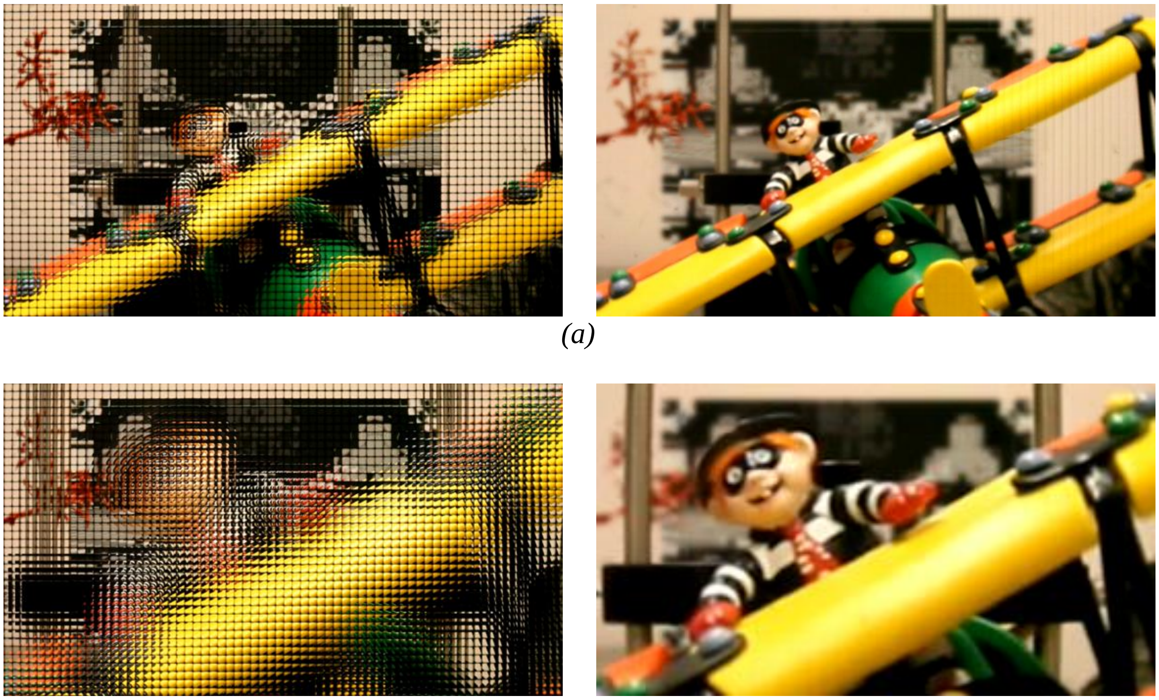
Figure A.7 Demichelis Spark LF sequence (1st frame): (a) LF content; and (b) Rendered central

Table A.3 Main characteristics of LF sequence Demichelis Spark

LF Sequence	*Demichelis Spark*
LF Resolution	2880×1620 (original) and 2850×1558 (rectified)
Frame Rate	25 Hz (150 frames)
Camera Setup	Focused (Plenoptic Camera 2.0)
MLA	Rectangular-arranged, semi-spherical lenses
Microlens	0.3 mm (pitch), 2.2 mm (focal length)
MI Resolution	38×38
Description	Natural studio scene with a presenter reading the script. Behind him, there is a static background. Camera moves linearly to the right.

A.2.3 Plane and Toy

Plane and Toy is shown in Figure A.8, and Table A.4 summarizes its main characteristics.



(a)

(b)

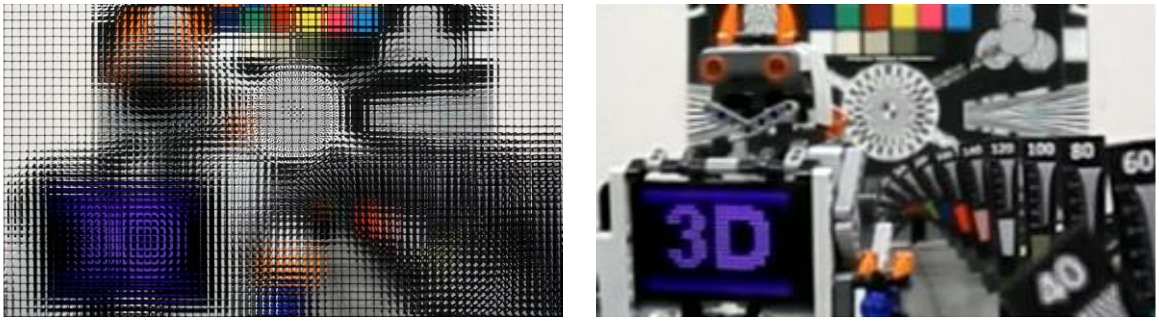
Figure A.8 Plane and Toy LF sequence: (a) LF content (left) and rendered central view (right) of frame number 23; and (b) LF content (left) and rendered central view (right) of frame number 123

Table A.4 Main characteristics of LF sequence Plane and Toy

LF Sequence	*Plane and Toy*
LF Resolution	1920×1088 (original) and 1904×1064 (rectified)
Frame Rate	25 Hz (250 frames)
Camera Setup	Focused (Plenoptic Camera 2.0)
MLA	Rectangular-arranged, squared-based lenses
Microlens	0.25 mm (pitch), 1.0 mm (focal length)
MI Resolution	~28×28 (integer approximation)
Description	Natural studio scene with a plane moving closer and nearer to the camera. Behind it, there is a static background.

A.2.4 Robot 3D

Robot 3D is shown in Figure A.9, and Table A.5 summarizes its main characteristics.



(a) *(b)*

Figure A.9 Robot test sequence (frame number 54): (a) LF content; and (b) Rendered central view

Table A.5 Main characteristics of LF sequence Robot 3D

LF Sequence	*Robot 3D*
LF Resolution	1920×1088 (original) and 1904×1064 (rectified)
Frame Rate	25 Hz (150 frames)
Camera Setup	Focused (Plenoptic Camera 2.0)
MLA	Rectangular-arranged, squared-based lenses
Microlens	0.25 mm (pitch), 1.0 mm (focal length)
MI Resolution	~28×28 (integer approximation)
Description	Natural studio scene with a robot moving around. Behind it, there is a static background with ISO and color charts.

Appendix B

Additional Results for Coding of Lytro Illum Light Field Images

This appendix presents additional RD performance results for the LFC solution with SS compensated prediction proposed in Chapter 4, in this case, considering LF images captured using a Lytro Illum LF camera. For this, the evaluation conditions are briefly presented in Section B.1 and, then, the experimental results are presented in Section B.2.

These additional results have been published in the following conference [56]:

- CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. 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, July 2016. p. 1–4.

B.1 Evaluation Conditions

To evaluate the performance of the LFC Bi-SS solution proposed in Chapter 4, the common test conditions defined in [106] are adopted, as summarized in the following.

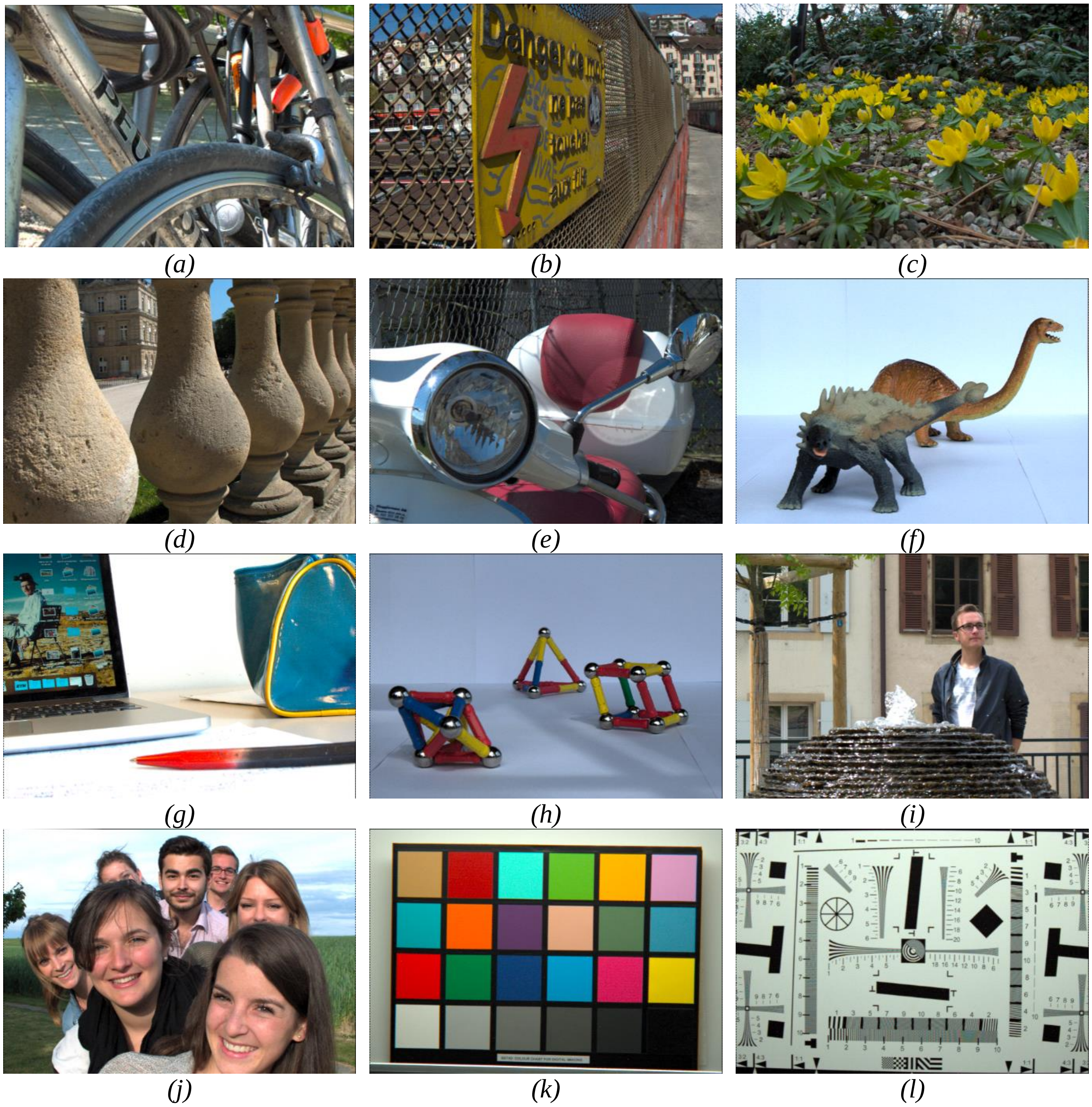
B.1.1 Test Images

Twelve LF test images provided by the ICME 2016 grand challenge in light field image compression [106] are used, as illustrated in Figure B.1. These LF images have been selected from the light field image dataset available in [218], and their main characteristics can be seen in Table A.1.

Each (raw) LF image was pre-processed before coding according to the processing flow defined in [106], and the input to the encoder is an LF image with Y’CbCr 4:2:0 color format. Please refer to [106, 222] for more details about the considered processing flow.

B.1.2 Test Conditions

The test conditions defined for the ICME 2016 grand challenge (as presented in the call for proposals document in [106]) are here adopted.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

(g) *(h)* *(i)*

(j) *(k)* *(l)*

Figure B.1 Central View rendered from the LF images (available in [218]*): (a) I01 – Bikes; (b) I02 – Danger_de_Mort; (c) I03 – Flowers; (d) I04 – Stone_Pillars_Outside; (e) I05 – Vespa; (f) I06 – Ankylosaurus_&_Diplodocus_1; (g) I07 – Desktop; (h) I08 – Magnets_1; (i) I09 – Fountain_&_Vincent_2; (j) I10 – Friends_1; (k) I11 – Color_Chart_1; and (l) I12 – ISO_Chart_12*

Each LF image with Y'CbCr 4:2:0 color format is encoded by the tested solutions for target Compression Ratio (CR) values of 10, 20, 40, and 100 (according to [106]). The target coding bits (corresponding to each target CR) are considered to be the maximum allowed amount of bits to encode the LF test images. This consideration also corresponds to the case where the target coding bits are used to define a restriction in the storage or channel capacity. Therefore, for each RD point, the QP value [2] of the encoder is adjusted to have the closest

Table B.1 Main characteristics of the Lytro Illum LF images (available in [218]*)*

LF Image	*I01*, *I02*, *I03*, *I04*, *I05*, *I06*, *I07*, *I08*, *I09*, *I10*, *I11*, and *I12*
LF Resolution	7728×5368
Camera Setup	Traditional LF Camera
MLA	Hexagonal-arranged, circular-based lenses
MI Resolution	~15×15 (integer approximation)

number of coding bits that is equal to or smaller than the target coding bits specified by the target CR.

B.1.3 Objective Quality Metrics

The objective quality evaluation was performed using the procedure outlined in [106], and the results are shown in terms of the mean YUV PSNR, $PSNR_{YUV_{mean}}$, and mean Y PSNR, $PSNR_{Y_{mean}}$, of all rendered viewpoint images. Details about these two objective metrics (i.e., $PSNR_{YUV_{mean}}$ and $PSNR_{YUV_{mean}}$) can be found in [106, 222].

B.1.4 Tested Solutions

The following three benchmark solutions are presented and compared against the proposed LFC Bi-SS solution:

1) **JPEG** – The reconstructed viewpoint images provided in [106] (anchors) are here used for comparison.
2) **HEVC** – The LF test images are encoded with the HEVC reference software version 14.0, using the Main Still Picture Profile [2].
3) **LFC Uni-SS** – The LF test images are encoded with the solution proposed in Chapter 4; however, in this case, only the uni-predicted SS candidate is available for the SS estimation and compensation (i.e., no bi-predicted SS candidate).

For both LFC Uni-SS and LFC Bi-SS solutions, a search range value of 128 is adopted for all LF images, and the full search algorithm with the HEVC quarter-pixel accuracy is used.

B.2 LFC Bi-SS RD Performance for Lytro Illum LF Image Coding

In this section, the LFC Bi-SS RD performance (using the BD metrics [210]) is presented in terms of $PSNR_{YUV_{mean}}$, in Table B.2, and in terms of $PSNR_{Y_{mean}}$, in Table B.3, for each LF image in Figure B.1.

Table B.2 BD results, in terms of $PSNR_{YUV_{mean}}$, for Lytro Illum LF image coding using the proposed LFC Bi-SS solution

LF Image	**JPEG**		**HEVC**		**LFC Uni-SS**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
I01	5.13	-66.57	1.06	-29.89	0.38	-12.70
I02	5.48	-70.56	0.68	-20.82	0.25	-8.60
I03	4.40	-58.06	0.23	-6.83	0.09	-2.91
I04	5.18	-62.25	0.25	-8.93	0.05	-1.76
I05	4.22	-71.75	0.69	-35.28	0.29	-19.53
I06	5.49	-79.12	1.52	-59.99	0.63	-42.55
I07	4.47	-68.64	0.37	-15.75	0.12	-5.79
I08	4.26	-72.60	0.73	-37.62	0.46	-29.04
I09	5.88	-78.74	1.54	-45.26	0.34	-13.50
I10	4.00	-72.66	0.15	-11.69	0.05	-5.33
I11	6.19	-85.78	1.74	-66.22	0.53	-34.10
I12	7.15	-81.36	1.61	-53.07	0.42	-22.19
Average	**5.15**	**-72.34**	**0.88**	**-32.61**	**0.30**	**-16.50**

Table B.3 BD results, in terms of $PSNR_{Y_{mean}}$, for Lytro Illum LF image coding using the proposed LFC Bi-SS solution

s	**JPEG**		**HEVC**		**LFC Uni-SS**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
I01	5.65	-65.67	1.21	-29.73	0.45	-13.08
I02	5.90	-69.22	0.79	-21.12	0.27	-8.06
I03	4.71	-55.18	0.25	-6.56	0.09	-2.49
I04	5.30	-59.83	0.32	-9.39	0.07	-2.10
I05	4.71	-70.43	0.76	-34.26	0.30	-18.21
I06	5.79	-77.41	1.71	-56.86	0.67	-39.14
I07	4.56	-68.32	0.37	-15.13	0.11	-4.97
I08	4.31	-69.83	0.80	-37.13	0.47	-27.78
I09	6.70	-77.11	1.82	-44.88	0.37	-12.66
I10	4.14	-71.12	0.17	-13.36	0.07	-7.23
I11	6.57	-85.17	**1.88**	**-66.97**	**0.56**	**-34.17**
I12	8.21	-80.46	1.80	-52.96	0.43	-20.89
Average	**5.55**	**-70.81**	**0.99**	**-32.36**	**0.32**	**-15.90**

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update*. 2015–2020 Cisco White Paper, 2016.

[2] SULLIVAN, Gary J., OHM, Jens-Rainer, HAN, Woo-Jin and WIEGAND, Thomas. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*. December 2012. Vol. 22, no. 12, p. 1649–1668.

[3] KATZMAIER, David. With a Bullet to the Head from Samsung, 3D TV is now deader than Ever. *CNET* [online]. 1 March 2016. Available from: https://www.cnet.com/news/3d-tv-is-now-more-dead-than-ever/

[4] CHARLTON, Alistair. 3D Television is Dead: Samsung and LG Cut Back on 3D TV Production. *International Business Times* [online]. 8 February 2016. Available from: http://www.ibtimes.co.uk/3d-television-dead-samsung-lg-cut-back-3d-tv-production-1542580

[5] AGGOUN, Amar, TSEKLEVES, Emmanuel, SWASH, Mohammad Rafiq, ZARPALAS, Dimitrios, DIMOU, Anastasios, DARAS, Petros, NUNES, Paulo and SOARES, Luís Ducla. Immersive 3D Holoscopic Video System. *IEEE Multimedia*. October 2013. Vol. 20, no. 1, p. 28–37. DOI 10.1109/MMUL.2012.42.

[6] Oculus Rift VR. [online]. 2016. [Accessed 2 November 2016]. Available from: https://www3.oculus.com/en-us/rift/

[7] SHEETZ, Michael. Virtual Reality is Finally for Real, and for You. *CNBC.com* [online]. 19 December 2015. Available from: http://www.cnbc.com/2015/12/18/virtual-reality-is-finally-for-real-and-for-you.html

[8] LIPPMANN, Gabriel. Épreuves Réversibles Donnant la Sensation du Relief. *Journal de Physique Théorique et Appliquée*. January 1908. Vol. 7, no. 1, p. 821–825.

[9] XIAO, Xiao, JAVIDI, Bahram, MARTINEZ-CORRAL, Manuel and STERN, Adrian. Advances in Three-Dimensional Integral Imaging: Sensing, Display, and Applications [Invited]. *Applied Optics*. February 2013. Vol. 52, no. 4, p. 546–560.

[10] GEORGIEV, Todor and LUMSDAINE, Andrew. Focused Plenoptic Camera and Rendering. *Journal of Electronic Imaging*. April 2010. Vol. 19, no. 2, p. 021106–021106. DOI 10.1117/1.3442712.

[11] RAYTRIX. Raytrix Website. [online]. 2012. [Accessed 7 July 2014]. Available from: http://www.raytrix.de/

[12] Lytro Inc. [online]. 2012. [Accessed 7 July 2016]. Available from: https://www.lytro.com/

[13] GEORGIEV, Todor, YU, Zhan, LUMSDAINE, Andrew and GOMA, Sergio. Lytro Camera Technology: Theory, Algorithms, Performance Analysis. In : *Proc. SPIE 8667, Multimedia Content and Mobile Devices*. Burlingame, CA, US, 7 March 2013. p. 86671J.

[14] ARAI, Jun, KAWAKITA, Masahiro, YAMASHITA, Takayuki, SASAKI, Hisayuki, MIURA, Masato, HIURA, Hitoshi, OKUI, Makoto and OKANO, Fumio. Integral Three-Dimensional Television with Video System Using Pixel-Offset Method. *Optics Express*. February 2013. Vol. 21, no. 3, p. 3474–3485. DOI 10.1364/OE.21.003474.

[15] ARAI, Jun. Integral Three-Dimensional Television. In : *2015 14th Workshop on Information Optics (WIO)*. Kyoto, Japan, June 2015. p. 1–3. ISBN 978-1-4673-7260-2.

[16] NHK STRL Science & Technology Research Laboratories. [online]. 2016. [Accessed 10 July 2016]. Available from: https://www.nhk.or.jp/strl/index-e.html

[17] BALOGH, Tibor, KOVACS, Peter Tamas and BARSI, Attila. Holovizio 3D Display System. In : *2007 3DTV Conference*. Kos Island, Greece, May 2007. p. 1–4. ISBN 978-1-4244-0721-7.

[18] HOLOGRAFIKA Holographic Display Technology. [online]. [Accessed 10 July 2016]. Available from: http://www.holografika.com/

[19] ISHIKAWA, Akio, PANAHPOUR TEHRANI, Mehrdad, NAITO, Sei, SAKAZAWA, Shigeyuki and KOIKE, Atsushi. Free Viewpoint Video Generation for Walk-Through Experience Using Image-Based Rendering. In : *Proceeding of the 16th ACM international conference on Multimedia - MM '08*. Vancouver, Canada, October 2008. p. 1007–1008. ISBN 9781605583037.

[20] Replay Technologies. [online]. 2016. [Accessed 16 November 2016]. Available from: http://replay-technologies.com/

[21] RAGHAVENDRA, R, RAJA, Kiran B and BUSCH, Christoph. Presentation Attack Detection for Face Recognition Using Light Field Camera. *IEEE Transactions on Image Processing*. March 2015. Vol. 24, no. 3, p. 1060–75. DOI 10.1109/TIP.2015.2395951.

[22] SHIN, Donghak, CHO, Myungjin and JAVIDI, Bahram. Three-Dimensional Optical Microscopy Using Axially Distributed Image Sensing. *Optics Letters*. November 2010. Vol. 35, no. 21, p. 3646. DOI 10.1364/OL.35.003646.

[23] WANG, Jingang, XIAO, Xiao, HUA, Hong and JAVIDI, Bahram. Augmented Reality 3D Displays With Micro Integral Imaging. *Journal of Display Technology*. November 2015. Vol. 11, no. 11, p. 889–893. DOI 10.1109/JDT.2014.2361147.

[24] LANMAN, Douglas and LUEBKE, David. Near-Eye Light Field Displays. *ACM SIGGRAPH 2013 Emerging Technologies - SIGGRAPH '13*. Anaheim, CA, US, July 2013. p. 1–1.

[25] VAN OOSTEROM, Peter, MARTINEZ-RUBI, Oscar, IVANOVA, Milena, HORHAMMER, Mike, GERINGER, Daniel, RAVADA, Siva, TIJSSEN, Theo, KODDE, Martin and GONÇALVES, Romulo. Massive Point Cloud Data Management: Design, Implementation and Execution of a Point Cloud Benchmark. *Computers & Graphics*. June 2015. Vol. 49, p. 92–125. DOI 10.1016/j.cag.2015.01.007.

[26] Mobile Mapping System - Mitsubishi Electric. [online]. [Accessed 3 November 2016]. Available from: http://www.mitsubishielectric.com/bu/mms/

[27] ScanLAB Project - Shipping Galleries. [online]. [Accessed 13 November 2016]. Available from: http://scanlabprojects.co.uk/projects/sciencemuseum

[28] Holoportation - Microsoft Research. [online]. 2016. [Accessed 12 November 2016]. Available from: https://www.microsoft.com/en-us/research/project/holoportation-3/

[29] vTime - The VR Sociable Network. [online]. [Accessed 20 November 2016]. Available from: https://vtime.net/

[30] Oculus Connect 3. [online]. 2016. [Accessed 23 November 2016]. Available from: https://www.oculusconnect.com/

[31] *JPEG Pleno - Scope, Use Cases and Requirements Ver1.6*. Chengdu, China : ISO/IEC JTC1/SC29/WG1 N73030, 2016.

[32] MIHAYLOVA, Emilia (ed.). *Holography - Basic Principles and Contemporary Applications*. INTECH, 2014. ISBN 9789535111177.

[33] EBRAHIMI, Touradj. *JPEG PLENO Abstract and Executive Summary*. Sydney, Australia : ISO/IEC JTC 1/SC 29/WG1 N6922, 2015.

[34] JPEG - JPEG Pleno. [online]. [Accessed 25 November 2016]. Available from: https://jpeg.org/jpegpleno/index.html

[35] *JPEG Pleno Call for Proposals on Light Field Coding (Draft Version 2)*. Chengdu, China : ISO/IEC JTC 1/ SC29/WG1 N73013, 2016.

[36] *Information technology - Generic coding of moving pictures and associated audio information: Video*. ITU-T Recommendation H.262 (02/12), 2012.

[37] *Stereoscopic Television MPEG-2 Multi-View Profile*. ITU-R Report BT.2017, 1998.

[38] WIEGAND, T., SULLIVAN, G.J., BJØNTEGAARD, G. and LUTHRA, A. Overview of the H.264/AVC Video Coding Standard. *IEEE Transactions on Circuits and Systems for Video Technology*. July 2003. Vol. 13, no. 7, p. 560–576.

[39] VETRO, A., WIEGAND, T. and SULLIVAN, G.J. Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard. *Proceedings of the IEEE*. April 2011. Vol. 99, no. 4, p. 626–642. DOI 10.1109/JPROC.2010.2098830.

[40] TECH, Gerhard, CHEN, Ying, MULLER, Karsten, OHM, Jens-Rainer, VETRO, Anthony and WANG, Ye-Kui. Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. January 2016. Vol. 26, no. 1, p. 35–49. DOI 10.1109/TCSVT.2015.2477935.

[41] TEHRANI, Mehrdad P., SHIMIZU, Shinya, LAFRUIT, Gauthier, SENOH, Takanori, FUJII, Toshiaki, VETRO, Anthony and TANIMOTO, Masayuki. *Use Cases and Requirements on Free-viewpoint Television (FTV)*. Geneva, Switzerland : ISO/IEC JTC1/SC29/WG11 MPEG N14104, 2013.

[42] *Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation*. Warsaw, Poland : ISO/IEC JTC1/SC29/WG11 MPEG2015/N15348, 2015.

[43] STANKIEWICZ, Olgierd, WEGNER, Krzysztof, SENOH, Takanori, LAFRUIT, Gauthier, BARONCINI, Vittorio and TANIMOTO, Masayuki. *Revised Summary of Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation*. Chengdu, China : ISO/IEC JTC1/SC29/WG11 MPEG2016/N16523, 2016.

[44] *Ad Hoc Groups Established at MPEG 114*. San Diego, CA, US : ISO/IEC JTC 1/SC 29/WG11 N15905, 2016.

[45] *List of AHGs Established at the 107th Meeting in San Jose, USA*. San Jose, CA, US : ISO/IEC JTC1/SC29/WG11 N14112, 2016.

[46] TULVAN, Christian, MEKURIA, Rufael, LI, Zhu and LASERRE, Sebastien. *Use Cases for Point Cloud Compression (PCC)*. Geneva, Switzerland : ISO/IEC JTC1/SC29/WG11 MPEG2015/ N16331, 2016.

[47] MEKURIA, Rufael, TULVAN, Christian and LI, Zhu. *Requirements for Point Cloud Compression*. Geneva, Switzerland : ISO/IEC JTC1/SC29/WG11 MPEG2016/N16330, 2016.

[48] *Draft Call for Proposals for Point Cloud Compression*. Chengdu, China : ISO/IEC JTC1/SC29/WG11 MPEG2014/N16538, 2016.

[49] HINDS, Arianne T., DOYEN, Didier and LAFRUIT, Gauthier. *Call for Light Field Test Material Including Plenoptic Cameras and Camera Arrays*. Chengdu, China : ISO/IEC JTC1/SC29/WG11 N16532, 2016.

[50] *Technical Report of the Joint Ad hoc Group for Digital Representations of Light/Sound Fields for Immersive Media Applications*. Geneva, Switzerland : ISO/IEC JTC1/SC29/WG1N72033, and ISO/IEC JTC1/SC29/WG11N16352, 2016.

[51] FLIERL, Markus and GIROD, Bernd. *Video Coding with Superimposed Motion-Compensated Signals - Applications to H.264 and Beyond*. 1. Springer US, 2004.

[52] SULLIVAN, G.J. and WIEGAND, T. Rate-Distortion Optimization for Video Compression. *IEEE Signal Processing Magazine*. November 1998. p. 74–90.

[53] CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. New HEVC Prediction Modes for 3D Holoscopic Video Coding. In : *2012 19th IEEE International Conference on Image Processing*. Orlando, FL, US, September 2012. p. 1325–1328. ISBN 978-1-4673-2533-2.

[54] CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. Influence of Self-Similarity on 3D Holoscopic Video Coding Performance. In : *Proc. of the 18th Brazilian symposium on Multimedia and the web - WebMedia '12*. São Paulo, Brazil, October 2012. p. 131–134. ISBN 978-1-4503-1706-1.

[55] AGOOUN, A., FATAH, Obaidulah Abdul, FERNANDEZ, Juan C J.C., CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Acquisition, Processing and Coding of 3D Holoscopic Content for Immersive Video Systems. In : *2013 3DTV Vision Beyond Depth (3DTV-CON)*. Aberdeen, Scotland, October 2013. p. 1–4. ISBN 978-1-4799-1369-5.

[56] CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. 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, July 2016. p. 1–4. ISBN 978-1-5090-1552-8.

[57] CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Light Field Image Coding with Jointly Estimated Self-Similarity Bi-Prediction. *Submitted to Signal Processing: Image Communication*.

[58] FLIERL, M., WIEGAND, T. and GIROD, B. A Locally Optimal Design Algorithm for Block-Based Multi-Hypothesis Motion-Compensated Prediction. In : *Proceedings DCC '98 Data Compression Conference*. Snowbird, UT, US, March 1998. p. 239–248. ISBN 0-8186-8406-2.

[59] CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. HEVC-Based 3D Holoscopic Video Coding using Self-Similarity Compensated Prediction. *Signal Processing: Image Communication*. March 2016. Vol. 42, p. 59–78. DOI 10.1016/j.image.2016.01.008.

[60] CONTI, Caroline, LINO, João, NUNES, Paulo, SOARES, Luís Ducla and CORREIA, Paulo Lobato. Spatial Prediction Based on Self-Similarity Compensation for 3D Holoscopic Image and Video Coding. In : *2011 18th IEEE International Conference on Image Processing*. Brussels, Belgium, September 2011. p. 961–964. ISBN 978-1-4577-1303-3.

[61] CONTI, Caroline, LINO, João, NUNES, Paulo, SOARES, Luís Ducla and CORREIA, Paulo Lobato. Improved Spatial Prediction for 3D Holoscopic Image and Video Coding. In : *19th European Signal Processing Conference (EUSIPCO 2011)*. Barcelona, Spain, August 2011. p. 378–382.

[62] CONTI, Caroline, LINO, João, NUNES, Paulo and SOARES, Luís Ducla. Spatial and Temporal Prediction Scheme for 3D Holoscopic Video Coding based on H.264/AVC. In : *2012 19th International Packet Video Workshop (PV)*. Munich, German, May 2012. p. 143–148. ISBN 978-1-4673-0301-9.

[63] CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Inter-Layer Prediction Scheme for Scalable 3-D Holoscopic Video Coding. *IEEE Signal Processing Letters*. August 2013. Vol. 20, no. 8, p. 819–822. DOI 10.1109/LSP.2013.2267234.

[64] CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Using Self-Similarity Compensation for Improving Inter-Layer Prediction in Scalable 3D Holoscopic Video Coding. In : *Proc. SPIE 8856 Applications of Digital Image Processing XXXVI*. San Diego, CA, US, September 2013. p. 88561K. ISBN 9780819497062.

[65] CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. Display Scalable 3D Holoscopic Video Coding. *IEEE Communications Society MMTC E-Letter*. May 2014. Vol. 9, no. 3, p. 12–15.

[66] CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. 3D Holoscopic Video Representation and Coding Technology. In : KONDOZ, Ahmet and DAGIUKLAS, Tasos (eds.), *Novel 3D Media Technologies*. New York, NY, US : Springer New York, 2015. p. 71–96. ISBN 978-1-4939-2025-9, 978-1-4939-2026-6.

[67] CONTI, Caroline, SOARES, Luís Ducla and NUNES, Paulo. Light Field Coding with Field of View Scalability for Flexible Interaction. *Submitted to IEEE Transactions on Multimedia*.

[68] CONTI, Caroline, NUNES, Paulo and SOARES, Luís Ducla. Impact of Packet Losses in Scalable 3D Holoscopic Video Coding. In : *Proc. SPIE Optics, Photonics, e Digital Technologies for Multimedia Applications III*. Brussels, Belgium, May 2014. p. 91380E. ISBN 9781628410860.

[69] CONTI, Caroline, KOVACS, Peter Tamas, BALOGH, Tibor, NUNES, Paulo and SOARES, Luís Ducla. Light-Field Video Coding using Geometry-Based Disparity Compensation. In : *2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. Budapest, Hungary, July 2014. p. 1–4. ISBN 978-1-4799-4758-4.

[70] LUCAS, Luís F. R., CONTI, Caroline, NUNES, Paulo, SOARES, Luís Ducla, RODRIGUES, Nuno M. M., PAGLIARI, Carla L., DA SILVA, Eduardo A.B. and DE FARIA, Sergio M. M. 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, September 2014. p. 11–15. ISBN 9780992862619.

[71] PEREIRA, Fernando, A. B. DA SILVA, Eduardo and LAFRUIT, Gauthier. Plenoptic Imaging: Representation and Processing. In : CHELLAPPA, R. and THEODORIDIS, S. (eds.). Academic Press Library in Signal Processing, 2016.

[72] GERSHUN, A. The Light Field. *Journal of Mathematics and Physics*. April 1939. Vol. 18, no. 1–4, p. 51–151. DOI 10.1002/sapm193918151.

[73] LEONARDO, Da Vinci. *The Notebooks of Leonardo Da Vinci*. Oxford University Press, 1988. ISBN 978-0486225722.

[74] BOGUSZ, A. Holoscopy and Holoscopic Principles. *Journal of Optics*. November 1989. Vol. 20, no. 6, p. 281–284. DOI 10.1088/0150-536X/20/6/005.

[75] ADELSON, Edward H. and BERGEN, James R. The Plenoptic Function and the Elements of Early Vision. *Computational Models of Visual Processing*. 1991. P. 3--20.

[76] GORTLER, Steven J., GRZESZCZUK, Radek, SZELISKI, Richard and COHEN, Michael F. The Lumigraph. In : *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96*. New Orleans, LA, US, October 1996. p. 43–54. ISBN 0-89791-746-4.

[77] LEVOY, Marc and HANRAHAN, Pat. Light Field Rendering. In : *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96*. New Orleans, LA, US, October 1996. p. 31–42. ISBN 0-89791-746-4.

[78] The Stanford Multi-Camera Array. [online]. [Accessed 20 August 2016]. Available from: https://graphics.stanford.edu/projects/array/

[79] NG, Ren. *Digital Light Field Photography*. Stanford, CA, US : Ph.D Thesis, Stanford University, 2006.

[80] LUMSDAINE, Andrew and GEORGIEV, Todor. The Focused Plenoptic Camera. In : *2009 IEEE International Conference on Computational Photography (ICCP)*. San Francisco, CA, US, April 2009. p. 1–8. ISBN 978-1-4244-4534-9.

[81] GEORGIEV, Todor. New Results on the Plenoptic 2.0 Camera. In : *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA, US, November 2009. p. 1243–1247. ISBN 978-1-4244-5825-7.

[82] GEORGEIV, Todor, ZHENG, Ke Colin, CURLESS, Brian, SALESIN, David, NAYAR, Shree and INTWALA, Chintan. Spatio-Angular Resolution Tradeoff in Integral Photography. In : *EGSR '06 Proceedings of the 17th Eurographics conference on Rendering Techniques*. Nicosia, Cyprus, June 2006. p. 263–272.

[83] LUMSDAINE, Andrew, GEORGIEV, Todor G. and CHUNEV, Georgi. Spatial Analysis of Discrete Plenoptic Sampling. In : *Proc. SPIE 8299, Digital Photography VIII*. Burlingame, CA, US, 22 January 2012. p. 829909.

[84] GEORGIEV, Todor and LUMSDAINE, Andrew. Depth of Field in Plenoptic Cameras. In : *Eurographics 2009*. Munich, German, March 2009.

[85] DANSEREAU, D.G. Donald G., PIZARRO, Oscar and WILLIAMS, Stefan B. S.B. Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras. In : *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, US, June 2013. p. 1027–1034. ISBN 978-0-7695-4989-7.

[86] BISHOP, Tom E. and FAVARO, Paolo. Plenoptic Depth Estimation from Multiple Aliased Views. In : *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. Kyoto, Japan, September 2009. p. 1622–1629. ISBN 978-1-4244-4442-7.

[87] CHEN, Jie and CHAU, Lap-Pui. A Fast Adaptive Guided Filtering Algorithm for Light Field Depth Interpolation. In : *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. June 2014. p. 2281–2284. ISBN 978-1-4799-3432-4.

[88] JEON, Hae-Gon, PARK, Jaesik, CHOE, Gyeongmin, PARK, Jinsun, BOK, Yunsu, TAI, Yu-Wing and KWEON, In So. Accurate depth map estimation from a lenslet light field camera. In : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. p. 1547–1555. ISBN 978-1-4673-6964-0.

[89] YU, Zhan, GUO, Xinqing, LING, Haibing, LUMSDAINE, Andrew and YU, Jingyi. Line Assisted Light Field Triangulation and Stereo Matching. In : *2013 IEEE International Conference on Computer Vision*. Sydney, Australia, December 2013. p. 2792–2799. ISBN 978-1-4799-2840-8.

[90] FLEISCHMANN, Oliver and KOCH, Reinhard. Lens-Based Depth Estimation for Multi-Focus Plenoptic Cameras. In : *Pattern Recognition*. Springer International Publishing, 2014. p. 410–420. Lecture Notes in Computer Science. ISBN 9783319117515.

[91] BOLLES, Robert C., BAKER, H. Harlyn and MARIMONT, David H. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*. March 1987. Vol. 1, no. 1, p. 7–55.

[92] OLSSON, Roger. Empirical Rate-Distortion Analysis of JPEG 2000 3D and H. 264/AVC Coded Integral Imaging Based 3D-Images. In : *2008 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*. Istanbul, Turkey, May 2008. p. 113–116. ISBN 978-1-4244-1760-5.

[93] OLSSON, Roger. *Synthesis, Coding, and Evaluation of 3D Images Based on Integral Imaging*. Sundsvall : Ph.D Thesis, Mid Sweden University, 2008.

[94] TOŠIĆ, Ivana and BERKNER, Kathrin. Light Field Scale-Depth Space Transform for Dense Depth Estimation. In : *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Columbus, OH, USA, June 2014. p. 441–448. ISBN 978-1-4799-4308-1.

[95] VAGHARSHAKYAN, Suren, BREGOVIC, Robert and GOTCHEV, Atanas. Image Based Rendering Technique via Sparse Representation in Shearlet Domain. In : *2015 IEEE International Conference on Image Processing (ICIP)*. Quebec City, Canada, September 2015. p. 1379–1383. ISBN 978-1-4799-8339-1.

[96] BRITES, Catarina, ASCENSO, Joao and PEREIRA, Fernando. Epipolar Plane Image Based Rendering for 3D Video Coding. In : *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*. Xiamen, China, October 2015. p. 1–6. ISBN 978-1-4673-7478-1.

[97] GEORGIEV, Todor and LUMSDAINE, Andrew. Rich Image Capture With Plenoptic Cameras. In : *2010 IEEE International Conference on Computational Photography (ICCP)*. Cambridge, MA, US, March 2010. p. 1–8. ISBN 978-1-4244-7022-8.

[98] LINO, João Filipe Oliveira. *2D Image Rendering for 3D Holoscopic Content using Disparity-Assisted Patch Blending*. Master Thesis, IST, University of Lisbon, 2013.

[99] DANSEREAU, Donald G., PIZARRO, Oscar and WILLIAMS, Stefan B. Linear Volumetric Focus for Light Field Cameras. *ACM Transactions on Graphics*. March 2015. Vol. 34, no. 2, p. 1–20. DOI 10.1145/2665074.

[100] GEORGIEV, Todor, CHUNEV, Georgi and LUMSDAINE, Andrew. Superresolution with the Focused Plenoptic Camera. In : *Proc. SPIE 7873, Computational Imaging IX*. San Francisco, CA, US, 10 February 2011. p. 78730X.

[101] EL-GHOROURY, Hussein S., CHUANG, Chih-Li and ALPASLAN, Zahir Y. Quantum Photonic Imager (QPI): A Novel Display Technology that Enables more than 3D Applications. *SID Symposium Digest of Technical Papers*. June 2015. Vol. 46, no. 1, p. 371–374. DOI 10.1002/sdtp.10255.

[102] Ostendo QPI. [online]. 2016. [Accessed 15 November 2016]. Available from: http://ostendo.com/media.html

[103] Microsoft HoloLens. [online]. [Accessed 5 November 2016]. Available from: https://www.microsoft.com/microsoft-hololens/en-us

[104] FORMAN, Matthew C., DAVIES, Neil A. and MCCORMICK, Malcolm. Objective Quality Measurement of Integral 3D Images. In : *Proc. SPIE 4660, Stereoscopic Displays and Virtual Reality Systems IX*. San Jose, CA, US : International Society for Optics and Photonics, May 2002. p. 155–162.

[105] SGOUROS, Nicholas, KONTAXAKIS, Ioannis and SANGRIOTIS, Manolis. Effect of Different Traversal Schemes in Integral Image Coding. *Applied Optics*. July 2008. Vol. 47, no. 19, p. D28. DOI 10.1364/AO.47.000D28.

[106] ŘEŘÁBEK, Martin, BRUYLANTS, Tim, EBRAHIMI, Touradj, PEREIRA, Fernando and SCHELKENS, Peter. *Call for Proposals and Evaluation Procedure*. Seattle, WA, US : ICME 2016 Grand Challenge: Light Field Image Compression, 2016.

[107] DANSEREAU, Donald. Light Field Toolbox v0.4. *MathWorks* [online]. 25 February 2015. [Accessed 10 February 2016]. Available from: http://www.mathworks.com/matlabcentral/fileexchange/49683

[108] *Methodology for the Subjective Assessment of the Quality of Television Pictures*. 01/2012. Recommendation ITU-R BT.500-13, 2012.

[109] VIOLA, Irene, ŘEŘÁBEK, Martin and EBRAHIMI, Touradj. A New Approach to Subjectively Assess Quality of Plenoptic Content. In : *Proc. SPIE 9971, Applications of Digital Image Processing XXXIX*. San Diego, CA, US, September 2016. p. 99710X.

[110] *Video Codec for Audiovisual Services at px64 kbits*. ITU-T Recommendation H.261 (03/93), 1993.

[111] *Information Technology - Digital Compression and Coding of Continuous-Tone Still Images - Requirements and Guidelines*. ITU-T Recommendation T.81 (09/92), 1992.

[112] *Information Technology - JPEG 2000 Image Coding System: Core Coding System*. ITU-T Recommendation T.800 (11/15), 2015.

[113] REHNA, V. J. and JEYA KUMAR, M. K. Hybrid Approaches to Image Coding: A Review. *International Journal of Advanced Computer Science and Applications*. 2011. Vol. 2, no. 7. DOI 10.14569/IJACSA.2011.020716.

[114] WIEN, Mathias. *High Efficiency Video Coding*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2015. Signals and Communication Technology. ISBN 978-3-662-44275-3.

[115] *Studio Encoding Parameters of Digital Television for Standard 4:3 and Wide Screen 16:9 Aspect Ratios*. Recommendation ITU-R BT.601-7 (03/2011), 2011.

[116] *Parameter Values for the HDTV Standards for Production and International Programme Exchange*. Recommendation ITU-R BT.709-6 (06/2015), 2015.

[117] *Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange*. Recommendation ITU-R BT.2020-2 (10/2015), 2015.

[118] RICHARDSON, Iain E. *The H.264 Advanced Video Compression Standard*. Wiley Publishing, 2010. ISBN 0470516925, 9780470516928.

[119] *Information Technology - Coding of Audio-Visual Objects - Part 2: Visual*. ISO/IEC 14496-2:2004, 2004.

[120] RICHARDSON, Iain E. G. *H.264 and MPEG-4 Video Compression : Video Coding for Next-Generation Multimedia*. Wiley, 2003. ISBN 9780470869604.

[121] ANTONINI, M., BARLAUD, M., MATHIEU, P. and DAUBECHIES, I. Image Coding Using Wavelet Transform. *IEEE Transactions on Image Processing*. April 1992. Vol. 1, no. 2, p. 205–220. DOI 10.1109/83.136597.

[122] ADAM, M.D. and KOSSENTNI, F. Reversible Integer-to-Integer Wavelet Transforms for Image Compression: Performance Evaluation and Analysis. *IEEE Transactions on Image Processing*. June 2000. Vol. 9, no. 6, p. 1010–1024. DOI 10.1109/83.846244.

[123] SZE, Vivienne, BUDAGAVI, Madhukar and SULLIVAN, Gary J. (eds.). *High Efficiency Video Coding (HEVC): Algorithms and Architectures*. Cham : Springer International Publishing, 2014. Integrated Circuits and Systems. ISBN 978-3-319-06894-7.

[124] TUNG NGUYEN and MARPE, D. Performance Analysis of HEVC-Based Intra Coding for Still Image Compression. In : *2012 Picture Coding Symposium*. May 2012. p. 233–236. ISBN 978-1-4577-2049-9.

[125] HANHART, Philippe, ŘEŘÁBEK, Martin, KORSHUNOV, Pavel and EBRAHIMI, Touradj. *Subjective Evaluation of HEVC Intra Coding for Still Image Compression*. Geneva, Switzerland : JCTVC-L0380, 2013.

[126] UGUR, Kemal and LAINEMA, Jani. *Updated Results on HEVC Still Picture Coding Performance*. Incheon, South Korea : JCTVC-M0041, 2013.

[127] LI, Yun. *Coding of Three-Dimensional Video Content*. Ph.D Thesis, Mid Sweden University, 2015.

[128] ALVES, Gustavo, PEREIRA, Fernando and DA SILVA, Eduardo A.B. Light Field Imaging Coding: Performance Assessment Methodology and Standards Benchmarking. In : *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. Seattle, WA, US, July 2016. p. 1–6. ISBN 978-1-5090-1552-8.

[129] SAXENA, Ankur and FERNANDES, Felix C. DCT/DST-Based Transform Coding for Intra Prediction in Image/Video Coding. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*. 1 October 2013. Vol. 22, no. 10, p. 3974–81. DOI 10.1109/TIP.2013.2265882.

[130] ASSUNÇÃO, Pedro, PINTO, Luís and FARIA, Sérgio. 3D Media Representation and Coding. In : *3D Future Internet Media*. New York, NY : Springer New York, 2014. p. 9–38.

[131] MERKLE, P., SMOLIC, A., MULLER, K. and WIEGAND, T. Efficient Prediction Structures for Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. November 2007. Vol. 17, no. 11, p. 1461–1473. DOI 10.1109/TCSVT.2007.903665.

[132] VETRO, Anthony and MÜLLER, Karsten. Depth-Based 3D Video Formats and Coding Technology. In : DUFAUX, Frédéric, PESQUET-POPESCU, Béatrice and CAGNAZZO, M (eds.), *Emerging Technologies for 3D Video*. Chichester, UK : John Wiley & Sons, Ltd, 2013. p. 139–161. ISBN 9781118583593.

[133] SCHARSTEIN, Daniel and SZELISKI, Richard. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*. April 2002. p. 7–42.

[134] FOIX, Sergi, ALENYA, Guillem and TORRAS, Carme. Lock-in Time-of-Flight (ToF) Cameras: A Survey. *IEEE Sensors Journal*. September 2011. Vol. 11, no. 9, p. 1917–1926. DOI 10.1109/JSEN.2010.2101060.

[135] FORMAN, Matthew C., AGGOUN, Amar and MCCORMICK, Malcolm. A Novel Coding Scheme for Full Parallax 3D-TV Pictures. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Munich, Germany, April 1997. p. 2945–2947.

[136] FORMAN, Matthew C. and AGGOUN, Amar. Compression of Full-Parallax Integral 3D-TV Image Data. In : *Proc. SPIE 3012, Stereoscopic Displays and Virtual Reality Systems IV*. San Jose, CA, US, May 1997. p. 222–226.

[137] FORMAN, Matthew C. and AGGOUN, Amar. Quantisation Strategies for 3D-DCT-Based Compression of Full Parallax 3D Images. In : *6th International Conference on Image Processing and its Applications*. Dublin, Ireland, July 1997. p. 32–35. ISBN 0 85296 692 X.

[138] FORMAN, Matthew Charles. *Compression of Integral Three-dimensional Television Pictures*. Ph.D Thesis, De Montfort University, 2000.

[139] AGGOUN, Amar. A 3D DCT Compression Algorithm For Omnidirectional Integral Images. In : *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. Toulouse, France, May 2006. p. II-517-II-520. ISBN 1-4244-0469-X.

[140] SGOUROS, Nicholas P., CHAIKALIS, Dionisis P., PAPAGEORGAS, Panagiotis G. and SANGRIOTIS, Manolis S. Omnidirectional Integral Photography Images Compression Using the 3D-DCT. In : *Digital Holography and Three-Dimensional Imaging*. Vancouver, Canada, June 2007. p. DTuA2. ISBN 1-55752-838-1.

[141] ZAHARIA, Ramona, AGGOUN, Amar and MCCORMICK, Malcolm. Compression of Full Parallax Colour Integral 3D TV Image Data Based on Sub-Sampling of Chrominance Components. In : *Proceedings of the Data Compression Conference*. Snowbird, UT, US, March 2001. p. 527.

[142] ZAHARIA, Ramona, AGGOUN, Amar and MCCORMICK, Malcolm. Adaptive 3D-DCT Compression Algorithm for Continuous Parallax 3D Integral Imaging. *Signal Processing: Image Communication*. March 2002. Vol. 17, no. 3, p. 231–242. DOI 10.1016/S0923-5965(01)00020-0.

[143] MEHANNA, A., AGGOUN, A., ABDULFATAH, O., SWASH, M. R. and TSEKLEVES, E. Adaptive 3D-DCT based Compression Algorithms for Integral Images. In : *2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. London, UK, June 2013. p. 1–5. ISBN 978-1-4673-6047-0.

[144] PEARSON, Karl. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine Series 6*. November 1901. Vol. 2, no. 11, p. 559–572. DOI 10.1080/14786440109462720.

[145] HOTELLING, Harold. Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology*. 1933. Vol. 24, no. 6, p. 417–441. DOI 10.1037/h0071325.

[146] SAYOOD, Khalid. *Introduction to Data Compression*. Morgan Kaufmann, 2012. ISBN 9780124157965.

[147] JANG, Ju-Seog, YEOM, Seokwon and JAVIDI, Bahram. Compression of Ray Information in Three-Dimensional Integral Imaging. *Optical Engineering*. December 2005. Vol. 44, no. 12, p. 127001. DOI 10.1117/1.2148947.

[148] LINDE, Yoseph, BUZO, Andrés and GRAY, Robert. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*. January 1980. Vol. 28, no. 1, p. 84–95. DOI 10.1109/TCOM.1980.1094577.

[149] KANG, Ho-Hyun, SHIN, Dong-Hak and KIM, Eun-Soo. Compression of Sub-Image-Transformed Elemental Images in Integral Imaging. In : *Digital Holography and Three-Dimensional Imaging*. Vancouver, Canada, June 2007. p. DTuA6. ISBN 1-55752-838-1.

[150] KANG, Ho-Hyun, SHIN, Dong-Hak and KIM, Eun-Soo. Compression Scheme of Sub-Images Using Karhunen-Loeve Transform in Three-Dimensional Integral Imaging. *Optics Communications*. July 2008. p. 3640–3647.

[151] AGGOUN, Amar. Compression of 3D Integral Images Using 3D Wavelet Transform. *Journal of Display Technology*. November 2011. Vol. 7, no. 11, p. 586–592. DOI 10.1109/JDT.2011.2159359.

[152] H. ZAYED, Hala, E. KISHK, Sherin and M. AHMED, Hosam. 3D Wavelets with SPIHT Coding for Integral Imaging Compression. *International Journal of Computer Science and Network Security*. January 2012. Vol. 12, no. 1, p. 126–133.

[153] TAUBMAN, D. High Performance Scalable Image Compression with EBCOT. *IEEE Transactions on Image Processing*. July 2000. Vol. 9, no. 7, p. 1158–1170. DOI 10.1109/83.847830.

[154] R. S. HIGA, R. F. L. CHAVEZ, R. B. LEITE, R. ARTHUR, Y. Iano. Plenoptic Image Compression Comparison between JPEG, JPEG2000 and SPITH. *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT)*. June 2013. Vol. 3, no. 6, p. 1–6.

[155] PERRA, Cristian. On the Coding of Plenoptic Raw Images. In : *2014 22nd Telecommunications Forum Telfor (TELFOR)*. November 2014. p. 850–853. ISBN 978-1-4799-6191-7.

[156] DUFAUX, Frederic, SULLIVAN, Gary and EBRAHIMI, Touradj. The JPEG XR Image Coding Standard [Standards in a Nutshell]. *IEEE Signal Processing Magazine*. November 2009. Vol. 26, no. 6, p. 195–199, 204–204. DOI 10.1109/MSP.2009.934187.

[157] *Information Technology - JPEG 2000 Image Coding System: Extensions for Three-Dimensional Data*. ITU-T Recommendation T.809 (05/11), 2011.

[158] ELHARAR, E., STERN, Adrian, HADAR, Ofer and JAVIDI, Bahram. A Hybrid Compression Method for Integral Images Using Discrete Wavelet Transform and Discrete Cosine Transform. *Journal of Display Technology*. September 2007. Vol. 3, no. 3, p. 321–325. DOI 10.1109/JDT.2007.900915.

[159] MAZRI, Meriem and AGGOUN, Amar. Compression of 3D Integral Images Using Wavelet Decomposition. In : *Proc. SPIE 5150, Visual Communications and Image Processing*. Lugano, Switzerland, June 2003. p. 1181–1192.

[160] AGGOUN, A. and MAZRI, M. Wavelet-Based Compression Algorithm for Still Omnidirectional 3D Integral Images. *Signal, Image and Video Processing*. June 2008. Vol. 2, no. 2, p. 141–153. DOI 10.1007/s11760-007-0044-1.

[161] KISHK, Sherin, AHMED, Hosam Eldin Mahmoud and HELMY, Hala. Integral Images Compression using Discrete Wavelets and PCA. *International Journal of Signal Processing, Image Processing and Pattern Recognition*. June 2011. Vol. 4, no. 2, p. 65–78.

[162] AGGOUN, A and TABIT, M. Data Compression of Integral Images for 3D TV. In : *2007 3DTV Conference*. Kos Island, Greece, May 2007. p. 1–4. ISBN 978-1-4244-0721-7.

[163] FORMAN, Matthew C., AGGOUN, Amar and MCCORMICK, Malcolm. Compression of Integral 3D TV Pictures. *Fifth International Conference on Image Processing and its Applications*. Heriot-Watt University, UK, August 1995. p. 584–588.

[164] SGOUROS, N.P., ANDREOU, A.G., SANGRIOTIS, M.S., PAPAGEORGAD, P.G., MAROULIS, D.M. and THEOFANOUS, N.G. Compression of IP Images for Autostereoscopic 3D Imaging Applications. In : *3rd International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003*. Rome, Italy : IEEE, September 2003. p. 223–227. ISBN 953-184-061-X.

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

[166] YEOM, Sekwon, STERN, Adrian and JAVIDI, Bahram. Compression of 3D Color Integral Images. *Optics Express*. 19 April 2004. Vol. 12, no. 8, p. 1632. DOI 10.1364/OPEX.12.001632.

[167] LEE, Ju-Han, YOO, Cheol-Hwa, KANG, Ho-Hyun and KIM, Eun-Soo. Compression scheme by use of motion-compensated residual image transformed from elemental image array in three-dimensional integral imaging. In : *Proc. SPIE 7864, Three-Dimensional Imaging, Interaction, and Measurement, 78640Y*. San Francisco, CA, US, January 2011. p. 78640Y.

[168] YOO, Cheol-Hwa, KANG, Ho-Hyun and KIM, Eun-Soo. Enhanced Compression of Integral Images by Combined use of Residual Images and MPEG-4 Algorithm in Three-Dimensional Integral Imaging. *Optics Communications*. September 2011. Vol. 284, no. 20, p. 4884–4893. DOI 10.1016/j.optcom.2011.06.020.

[169] KANG, Ho-Hyun, LEE, Ju-Han and KIM, Eun-Soo. Enhanced Compression Rate of Integral Images by using Motion-Compensated Residual Images in Three-Dimensional Integral-Imaging. *Optics Express*. February 2012. Vol. 20, no. 5, p. 5440–5459. DOI 10.1364/OE.20.005440.

[170] LEE, Hyoung-Woo, LEE, Ju-Han, KANG, Ho-Hyun and KIM, Eun-Soo. Compression Enhancement using the Hybrid Motion Estimation in Sub-Image Array Transformed from Elemental Image Array in Three-Dimensional Integral Image. In : *Proc. SPIE 8498, Optics and Photonics for Information Processing VI, 849804*. San Diego, CA, US, October 2012. p. 849804.

[171] OLSSON, Roger, SJOSTROM, Marten and XU, Youzhi. A Combined Pre-Processing and H.264-Compression Scheme for 3D Integral Images. In : *2006 International Conference on Image Processing*. Atlanta, GA, US, 2006. p. 513–516. ISBN 1-4244-0480-0.

[172] OLSSON, Roger, SJÖSTRÖM, Mårten and XU, Youzhi. Evaluation of a Combined Pre-Processing and H.264-Compression Scheme for 3D Integral Images. In : *Proc. SPIE 6508, Visual Communications and Image Processing*. San Jose, CA, US, January 2007. p. 65082C.

[173] KANG, Ho-Hyun, SHIN, Dong-Hak and KIM, Eun-Soo. Efficient Compression of Motion-Compensated Sub-images with Karhunen–Loeve Transform in Three-dimensional Integral Imaging. *Optics Communications*. March 2010. Vol. 283, no. 6, p. 920–928. DOI 10.1016/j.optcom.2009.11.033.

[174] MATTOCCIA, Stefano, TOMBARI, Federico and DI STEFANO, Luigi. Fast Full-Search Equivalent Template Matching by Enhanced Bounded Correlation. *IEEE Transactions on Image Processing*. April 2008. Vol. 17, no. 4, p. 528–538. DOI 10.1109/TIP.2008.919362.

[175] VIEIRA, Alexandre, DUARTE, Helder, PERRA, Cristian, TAVORA, Luis and ASSUNCAO, Pedro. 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, November 2015. p. 494–497. ISBN 978-1-4799-8636-1.

[176] BOSSEN, F. *Common HM Test Conditions and Software Reference Configurations*. Geneva, Switzerland : JCTVC-L1100, 2013.

[177] FECKER, U and KAUP, A. H.264/AVC-Compatible Coding of Dynamic Light Fields Using Transposed Picture Ordering. *Signal Processing Conference, 2005 13th European*. Antalya, Turkey, September 2005. p. 1–4.

[178] ADEDOYIN, S., FERNANDO, W.A.C. and AGGOUN, A. A Joint Motion & Disparity Motion Estimation Technique for 3D Integral Video Compression Using Evolutionary Strategy. *IEEE Transactions on Consumer Electronics*. July 2007. p. 732–739.

[179] ADEDOYIN, S., FERNANDO, W.A.C., AGGOUN, A. and WEERAKKODY, W.A.R.J. An ES Based Effecient Motion Estimation Technique for 3D Integral Video Compression. In : *2007 IEEE International Conference on Image Processing*. San Antonio, TX, US, 2007. p. III-393-III-396. ISBN 978-1-4244-1436-9.

[180] ADEDOYIN, S., FERNANDO, W. A C, AGGOUN, A and KONDOZ, K. M. Motion and Disparity Estimation with Self Adapted Evolutionary Strategy in 3D Video Coding. *IEEE Transactions on Consumer Electronics*. November 2007. Vol. 53, no. 4, p. 1768–1775. DOI 10.1109/TCE.2007.4429282.

[181] DICK, Julien, ALMEIDA, Hugo, SOARES, Luís Ducla and NUNES, Paulo. 3D Holoscopic Video Coding Using MVC. In : *2011 IEEE EUROCON - International Conference on Computer as a Tool*. Lisbon, Portugal, April 2011. p. 1–4. ISBN 978-1-4244-7487-5.

[182] SHI, Shasha, GIOIA, Patrick and MADEC, Gerard. Efficient Compression Method for Integral Images using Multi-View Video Coding. In : *2011 18th IEEE International Conference on Image Processing*. Brussels, September 2011. p. 137–140. ISBN 978-1-4577-1303-3.

[183] WEI, Jian, WANG, Shigang, ZHAO, Yan and JIN, Fushou. Hierarchical Prediction Structure for Subimage Coding and Multithreaded Parallel Implementation in Integral Imaging. *Applied Optics*. 20 April 2011. Vol. 50, no. 12, p. 1707. DOI 10.1364/AO.50.001707.

[184] WANG, Gengkun, XIANG, Wei, PICKERING, Mark and CHEN, Chang Wen. Light Field Multi-View Video Coding With Two-Directional Parallel Inter-View Prediction. *IEEE Transactions on Image Processing*. November 2016. Vol. 25, no. 11, p. 5104–5117. DOI 10.1109/TIP.2016.2603602.

[185] DRICOT, Antoine, JUNG, Joel, CAGNAZZO, Marco, PESQUET, Béatrice and DUFAUX, Frédéric. Full Parallax Super Multi-View Video Coding. In : *2014 IEEE International Conference on Image Processing (ICIP)*. Paris, France, October 2014. p. 135–139. ISBN 978-1-4799-5751-4.

[186] YU, S.-L and CHRYSAFIS, C. *New Intra Prediction Using Intra-Macroblock Motion Compensation*. Fairfax, VA, US : JVT-C151, 2002.

[187] SIU-LEONG YU, Christos Chrysafis. Intra-Prediction Using Intra-Macroblock Motion Compensation. US7120196 B2. 2006. US : Google Patents.

[188] TÜRKAN, Mehmet and GUILLEMOT, Christine. Image Prediction Based on Neighbor-Embedding Methods. *IEEE Transactions on Image Processing*. April 2012. Vol. 21, no. 4, p. 1885–98. DOI 10.1109/TIP.2011.2170700.

[189] LI, Yun, SJOSTROM, Marten, OLSSON, Roger and JENNEHAG, Ulf. Coding of Focused Plenoptic Contents by Displacement Intra Prediction. *IEEE Transactions on Circuits and Systems for Video Technology*. July 2016. Vol. 26, no. 7, p. 1308–1319. DOI 10.1109/TCSVT.2015.2450333.

[190] XU, Jizheng, JOSHI, Rajan and COHEN, Robert A. Overview of the Emerging HEVC Screen Content Coding Extension. *IEEE Transactions on Circuits and Systems for Video Technology*. January 2016. Vol. 26, no. 1, p. 50–62. DOI 10.1109/TCSVT.2015.2478706.

[191] FLYNN, David, MARPE, Detlev, NACCARI, Matteo, NGUYEN, Tung, ROSEWARNE, Chris, SHARMAN, Karl, SOLE, Joel and XU, Jizheng. Overview of the Range Extensions for the HEVC Standard: Tools, Profiles, and Performance. *IEEE Transactions on Circuits and Systems for Video Technology*. January 2016. Vol. 26, no. 1, p. 4–19. DOI 10.1109/TCSVT.2015.2478707.

[192] BUDAGAVI, Madhukar and KWON, Do-Kyoung. Intra Motion Compensation and Entropy Coding Improvements for HEVC Screen Content Coding. In : *2013 Picture Coding Symposium (PCS)*. San Jose, CA, US, December 2013. p. 365–368. ISBN 978-1-4799-0294-1.

[193] ROSEWARNE, C., SHARMAN, K., NACCARI, M. and SULLIVAN, G. *HEVC Range Extensions Test Model 6 Encoder Description*. San Jose, CA, US : JCTVC-P1013, 2014.

[194] KWON, Do-Kyoung and BUDAGAVI, Madhukar. Fast Intra Block Copy (IntraBC) Search for HEVC Screen Content Coding. In : *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. Melbourne, Australia, June 2014. p. 9–12. ISBN 978-1-4799-3432-4.

[195] JOSHI, Rajan, XU, Jizheng, COHEN, Robert, LIU, Shan, MA, Zhan and YE, Yan. *Screen Content Coding Test Model 1 (SCM 1)*. Valencia, Spain : JCTVC-Q1014, 2014.

[196] LEE, Daniel D. and SEUNG, H. Sebastian. Algorithms for Non-Negative Matrix Factorization. In : *Proc. of NIPS*. Denver, CO, US, 2000. p. 556–562.

[197] ROWEIS, Sam T. and SAUL, Lawrence K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*. December 2000. Vol. 290, no. 5500, p. 2323–6. DOI 10.1126/science.290.5500.2323.

[198] TAN, Thiow, BOON, Choong and SUZUKI, Yoshinori. Intra Prediction by Template Matching. In : *2006 International Conference on Image Processing*. Atlanta, GA, US, 2006. p. 1693–1696. ISBN 1-4244-0480-0.

[199] LIU, Deyang, AN, Ping, MA, Ran and SHEN, Liquan. Disparity Compensation Based 3D Holoscopic Image Coding Using HEVC. In : *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. Chengdu, China, July 2015. p. 201–205. ISBN 978-1-4799-1948-2.

[200] LIU, Deyang, AN, Ping, MA, Ran, YANG, Chao and SHEN, Liquan. 3D Holoscopic Image Coding Scheme Using HEVC with Gaussian Process Regression. *Signal Processing: Image Communication*. September 2016. Vol. 47, p. 438–451. DOI 10.1016/j.image.2016.08.004.

[201] PIAO, Yan and YAN, Xiaoyuan. Sub-Sampling Elemental Images for Integral Imaging Compression. In : *2010 International Conference on Audio, Language and Image Processing*. Shanghai, China, November 2010. p. 1164–1168. ISBN 978-1-4244-5856-1.

[202] YAN, P. and XIANYUAN, Y. Integral Image Compression Based on Optical Characteristic. *IET Computer Vision*. May 2011. Vol. 5, no. 3, p. 164. DOI 10.1049/iet-cvi.2010.0031.

[203] ZHENG, Wanying and PIAO, Yan. Research on Integral (3D) Image Compression Technology Based on Neural Network. In : *2012 5th International Congress on Image and Signal Processing*. Chongqing, Sichuan, China, October 2012. p. 382–386. ISBN 978-1-4673-0964-6.

[204] CHOUDHURY, Chandrajit and CHAUDHURI, Subhasis. 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, December 2014. p. 1–6. ISBN 9781450330619.

[205] GRAZIOSI, Danillo B., ALPASLAN, Zahir Y. and EL-GHOROURY, Hussein S. Depth Assisted Compression of Full Parallax Light Fields. In : *Proc. SPIE 9391, Stereoscopic Displays and Applications XXVI*. San Francisco, CA, US, March 2015. p. 93910Y.

[206] LI, Yun, SJÖSTRÖM, Mårten, OLSSON, Roger and JENNEHAG, Ulf. Scalable Coding of Plenoptic Images by Using a Sparse Set and Disparities. *IEEE Transactions on Image Processing*. January 2016. Vol. 25, no. 1, p. 80–91. DOI 10.1109/TIP.2015.2498406.

[207] DRICOT, A., JUNG, J., CAGNAZZO, M., PESQUET, B. and DUFAUX, F. Integral Images Compression Scheme Based On View Extraction. In : *2015 23rd European Signal Processing Conference (EUSIPCO)*. Nice, France, August 2015. p. 101–105. ISBN 978-0-9928-6263-3.

[208] DRICOT, Antoine, JUNG, Joel, CAGNAZZO, Marco, PESQUET, Béatrice and DUFAUX, Frédéric. Improved Integral Images Compression Based on Multi-View Extraction. In : *Proc. SPIE 9971, Applications of Digital Image Processing XXXIX*. San Diego, CA, US, September 2016. p. 99710L.

[209] IL-KOO KIM, MCCANN, Ken, SUGIMOTO, Kazuo, BROSS, Benjamin, HAN, Woo-Jin and SULLIVAN, Gary. *High Efficiency Video Coding (HEVC) Test Model 14 (HM14) Encoder Description*. San José, CA, US : JCTVC-P1002, 2014.

[210] BJØNTEGAARD, Gisle. *Calculation of Average PSNR Differences between RD Curves*. Austin, TX, US : VCEG-M33, 2001.

[211] GIROD, Bernd. Efficiency Analysis of Multihypothesis Motion-Compensated Prediction for Video Coding. *IEEE Transactions on Image Processing*. January 2000. Vol. 9, no. 2, p. 173–83. DOI 10.1109/83.821595.

[212] RAMANATHAN, P. and GIROD, B. Rate-Distortion Analysis of Random Access for Compressed Light Fields. In : *2004 International Conference on Image Processing, ICIP '04.* Nanyang, Singapore, 2004. p. 2463–2466. ISBN 0-7803-8554-3.

[213] *White Paper on State of the Art in compression and transmission of 3D Video*. Geneva, Switzerland : ISO/IEC JTC1/SC29/WG11 N13364, 2013.

[214] MV-HEVC Reference Software HTM-12.0. [online]. [Accessed 22 December 2014]. Available from: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-12.0/

[215] CRIMINISI, A., PEREZ, P. and TOYAMA, K. Region Filling and Object Removal by Exemplar-Based Image Inpainting. *IEEE Transactions on Image Processing*. September 2004. Vol. 13, no. 9, p. 1200–1212. DOI 10.1109/TIP.2004.833105.

[216] BOSSEN, F., BROSS, B., SUHRING, K. and FLYNN, D. HEVC Complexity and Implementation Analysis. *IEEE Transactions on Circuits and Systems for Video Technology*. December 2012. Vol. 22, no. 12, p. 1685–1696.

[217] BUYSSENS, Pierre, DAISY, Maxime, TSCHUMPERLE, David and LEZORAY, Olivier. Exemplar-Based Inpainting: Technical Review and New Heuristics for Better Geometric Reconstructions. *IEEE Transactions on Image Processing*. March 2015. Vol. 24, no. 6, p. 1809–1824. DOI 10.1109/TIP.2015.2411437.

[218] ŘEŘÁBEK, Martin and EBRAHIMI, Touradj. New Light Field Image Dataset. In : *8th International Conference on Quality of Multimedia Experience (QoMEX)*. Lisbon, Portugal, 2016.

[219] PAUDYAL, Pradip, OLSSON, Roger, SJÖSTRÖM, Mårten, BATTISTI, Federica and CARLI, Marco. SMART: A Light Field Image Quality Dataset. In : *Proceedings of the 7th International Conference on Multimedia Systems - MMSys '16*. New York, NY, US, May 2016. p. 1–6. ISBN 9781450342971.

[220] LV, Hao, WANG, Ronggang, XIE, Xiaodong, JIA, Huizhu and GAO, Wen. A Comparison of Fractional-Pel Interpolation Filters in HEVC and H.264/AVC. In : . November 2012. p. 1–6.

[221] GEOGIEV, Todor. Todor Georgiev Gallery of Light Field Data. [online]. [Accessed 17 September 2016]. Available from: http://www.tgeorgiev.net/Gallery/

[222] VIOLA, Irene, ŘEŘÁBEK, Martin, BRUYLANTS, Tim, SCHELKENS, Peter, PEREIRA, Fernando and EBRAHIMI, Touradj. Objective and Subjective Evaluation of Light Field Image Compression Algorithms. In : *(To appear on) 32nd Picture Coding Symposium*. Nuremberg, Germany, 2016.