Compression and visual quality assessment for light field contents

Thèse N° 7185

Présentée le 26 avril 2019

à la Faculté des sciences et techniques de l'ingénieur
Groupe Ebrahimi
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Irene VIOLA

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury
Prof. T. Ebrahimi, directeur de thèse
Prof. E. A. B. da Silva, rapporteur
Prof. F. M. B. Pereira, rapporteur
Dr J.-M. Vesin, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

“Perfection is achieved not when there is nothing more to add,
but when there is nothing left to take away.”
— Antoine de Saint-Exupéry

Acknowledgements

The work of this thesis would not be possible without the help of all of the people who have supported and encouraged me. I would like to start by thanking Prof. Touradj Ebrahimi for supporting my PhD studies. I am beyond grateful for the endless opportunities he provided to help me grow and flourish as a researcher.

Besides my advisor, I would also like to thank the experts who served in my thesis committee: Prof. Pascal Frossard, Prof. Eduardo da Silva, Prof. Fernando Pereira and Dr. Jean-Marc Vesin. I deeply appreciate the time you spent to read and review my (admittedly long) manuscript, and I treasure the fruitful comments you have given me during my defense.

I would like to extend my gratitute to Prof. Toshiaki Fujii, for welcoming me in his laboratory in Nagoya University: it was an experience that has taught me a lot, both professionally and humanely, and I will cherish the memories forever. Thank you for the kind supervision, for the celebrations, and for the fun Italian-Japanese translations.

I had the privilege of sharing an office and more than three years of my life with wonderful collegues. First and foremost I would like to thank Martin, for the endless support during the early stages of my research: I don't know what would have become of me without your gentle help, but I am sure I would be a much worse researcher. To Lin and Ashu: the time I had to share with you might have been shorter, but it was full of fun and laughter, and I thank you for that. To Anne-Flore, my unofficial post-doc: thank you for the interesting discussions, both professional and not, the nice vacations, the silly singing, and all the great times we shared. To Eugene, thanks for the unusual points of view, which have sparked many engaging conversations. To Pinar, thank you for all the crazy times, for the tears and the laughter, the inside jokes, the stern conversations, the hugs and the kisses, and all the endless (and entertaining!) drama. To Vaggelis, thank you for being my rock.

I got to meet a lot of amazing people during these years, and for that I am beyond grateful. To the guys and girls with whom I got to share the joy and terror of first starting a PhD (Hermina, Helena, Patricia, Oana, Lana, George, Marios, Lefteris, Dusan, Kirill): I couldn't have asked for better people to live all these amazing and terrible times together. To Beril, Tolis, Melivoia, and all the guys and girls of the corridor: thank you for the loud lunches, the interesting conversations, and all the fun times we shared. To my flatmates, old and new (Enrica, Hermina, Miranda): no matter how hard the day was, you always made it a joy to come back home.

A special thanks goes to my friends, who have supported me throughout this long journey: to Cinzia and Celeste, for the 10+ years of friendship and for always being there for me, to Sylvie

and Davina, the best surrogate family I could ever ask for, to the ladies of the Twitterverse (Liv, Lalla, Alex, Lorenza), to my friends in Sicily, Turin and across the globe.
Finally, I would like to thank my family, for the endless support they provided during these years. Thank you for always believing in me, for encouraging my dreams, for supporting my every decision, no matter how far away it was bringing me from you. This work couldn't have been completed without you.

Abstract

Since its invention in the 19th century, photography has allowed to create durable images of the world around us by capturing the intensity of light that flows through a scene, first analogically by using light-sensitive material, and then, with the advent of electronic image sensors, digitally. However, one main limitation of both analog and digital photography lays in its inability to capture any information about the direction of light rays. Through traditional photography, each Three-dimensional scene is projected onto a 2D plane; consequently, no information about the position of the 3D objects in space is retained.

Light field photography aims at overcoming these limitations by recording the direction of light along with its intensity. In the past, several acquisition technologies have been presented to properly capture light field information, and portable devices have been commercialized to the general public. However, a considerably larger volume of data is generated when compared to traditional photography. Thus, new solutions must be designed to face the challenges light field photography poses in terms of storage, representation and visualization of the acquired data. In particular, new and efficient compression algorithms are needed to sensibly reduce the amount of data that needs to be stored and transmitted, while maintaining an adequate level of perceptual quality.

In designing new solutions to address the unique challenges posed by light field photography, one cannot forgo the importance of having reliable, reproducible means of evaluating their performance, especially in relation to the scenario in which they will be consumed. To that end, subjective assessment of visual quality is of paramount importance to evaluate the impact of compression, representation, and rendering models on user experience. Yet, the standardized methodologies that are commonly used to evaluate the visual quality of traditional media content, such as images and videos, are not equipped to tackle the challenges posed by light field photography. New subjective methodologies must be tailored for the new possibilities this new type of imaging offers in terms of rendering and visual experience.

In this work, we address the aforementioned problems by both designing new methodologies for visual quality evaluation of light field contents, and outlining a new compression solution to efficiently reduce the amount of data that needs to be transmitted and stored. We first analyse how traditional methodologies for subjective evaluation of multimedia contents can be adapted to suit light field data, and we propose new methodologies to reliably assess the visual quality while maintaining user engagement. Furthermore, we study how user behavior is affected by the visual quality of the data. We employ subjective quality assessment to compare several state-of-the-art solutions in light field coding, in order to find the most promising

approaches to minimize the volume of data without compromising on the perceptual quality. To that means, we define and inspect several coding approaches for light field compression, and we investigate the impact of color subsampling on the final rendered content. Lastly, we propose a new coding approach to perform light field compression, showing significant improvement with respect to the state of the art.

Keywords: Light field photography, immersive media, subjective quality evaluation, objective quality evaluation, interactive assessment, passive assessment, light field displays, light field rendering, light field compression, graph-based compression

Sommario

Fin dalla sua invenzione nel diciannovesimo secolo, la fotografia ha permesso di creare immagini durature del mondo intorno a noi. Ciò è ottenuto catturando l'intensità della luca che passa attraverso una scena, prima analogicamente usando materiali fotosensibili, e poi, con l'avvento di sensori ottici elettronici, digitalmente. Tuttavia, una delle maggiori limitazioni della fotografia analogica e digitale si riscontra nella sua incapacità di registrare alcuna informazione sulla direzione dei raggi luminosi. Usando metodi fotografici tradizionali, le scene tridimensionali sono proiettate su un piano bidimensionale; consequentemente, la posizione degli oggetti 3D nello spazio non è catturata.

L'obiettivo della fotografia light field è di superare queste limitazioni, attraverso l'acquisizione della direzione della luce, e non solo della sua intensità. In passato, diverse tecnologie per l'acquisizione di light field sono state presentate, e dispositivi portatili sono stati commercializzati per il pubblico generale. Tuttavia, il volume di dati generato è notevolmente più grande, se paragonato alla fotografia tradizionale.

Per questo motivo, nuove soluzioni devono essere progettate per affrontare le sfide poste dalla fotografia light field in termini di archiviazione, rappresentazione e visualizzazione dei dati acquisiti. In particolare, nuovi algoritmi di compressione più efficienti devono essere pensati per ridurre notevolmente la quantità di informazione che dev'essere memorizzata e trasmessa, al contempo mantenendo un livello adeguato di qualità percettiva.

Nel progettare nuove soluzioni per affrontare le sfide uniche che la fotografia light field pone, non si può dimenticare l'importanza di avere dei metodi affidabili e riproducibili per valutare le loro prestazioni, specialmente in relazione allo scenario in cui verranno fruiti. Per questo scopo, la valutazione soggettiva della qualità visiva è importantissima per stimare l'impatto dei modelli di compressione, rappresentazione e visualizzazione sull'esperienza dell'utente. Ciò nonostante, le metodologie standardizzate comunemente usate per stimare la qualità visiva dei contenuti multimediali tradizionali, come immagini e video, non sono adatte ad affrontare le sfide poste dalla tecnologia light field. Nuove metodologie soggettive devono essere progettate su misura per le nuove possibilità offerte da questo nuovo tipo di tecnologia in termini di visualizzazione ed esperienza visiva.

In questo lavoro affrontiamo i problemi sopra citati in due modi: progettando nuove metodologie per valutare la qualità visiva dei contenuti light field, e proponendo una nuova soluzione per la compressione di light field che riduce il volume di dati necessario per la trasmissione e memorizzazione. Per prima cosa, analizziamo come le metodologie tradizionali per la valutazione soggettiva dei contenuti multimediali possono essere adattate per i contenuti light field,

e proponiamo nuove metodologie per stimare affidabilmente la qualità visiva mantenendo il coinvolgimento degli utenti. In più, studiamo come il comportamento degli utenti può essere influenzato dalla qualità visiva dei dati. Usiamo metodi per la valutazione della qualità soggettiva per mettere a confronto diverse soluzioni all'avanguardia nella compressione di light field, per trovare l'approccio più promettente per minimizzare il volume di dati senza compromettere la qualità perceptita. Per aiutarci nel nostro intento, definiamo e analizziamo diversi approcci per la compressione di light field, e indaghiamo sull'impatto del sottocampionamento della crominanza sul prodotto finale visualizzato. Infine, proponiamo un nuovo approccio per la compressione di light field, e mostriamo dei miglioramenti notevoli nei confronti degli ultimi approcci all'avanguardia.

Parole chiave: Fotografia light field, contenuti multimediali immersivi, valutazione della qualità soggettiva, valutazione della qualità oggettiva, valutazione interattiva, valutazione passiva, display light field, rendering light field, compressione light field, compressione basata su grafi

Contents

List of Figures

List of Tables

List of Acronyms

2D Two-dimensional. vii, 1, 3, 4, 9, 15, 17, 22, 48, 65, 66, 77, 120, 124, 163, 171, 181, 182

3D Three-dimensional. vii, ix, 7, 10, 11, 15, 17, 21, 34, 65, 129, 167, 181

ACR Absolute Category Rating. 15, 16, 23, 47, 48

ANOVA Analysis of variance. xxi, 23, 27–33, 40, 43, 101, 102

AR Augmented Reality. 11

AVC Advanced Video Coding. xxi, 13, 38, 39

bpp bit per pixel. 24, 37, 51, 67, 83, 95, 97, 116, 121, 123, 124, 131, 151, 153, 172, 173

CFP Call For Proposals. 145

CIs Confidence Intervals. xv–xviii, 3, 27, 40, 44, 45, 48, 53, 56, 59, 62, 69, 70, 74, 75, 90, 101, 102, 105, 108, 118, 121, 122, 134, 136, 137, 139, 162, 163, 171

CNN Convolutional Neural Network. 12, 13

DCT Discrete Cosine Transform. 13, 14, 157

DOF Depth Of Field. 2, 15, 23, 25, 34, 36, 51

DQR Degradation Quality Rating. 15, 47

DSCQS Double Stimulus Continuous Quality Scale. 25, 26, 166

DSIS Double Stimulus Impairment Scale. xii, xxii, 37, 52, 67, 69, 70, 73, 75, 77, 78, 133, 136, 163, 176, 178

EPFL École Polytechnique Fédérale de Lausanne. 68–70, 74, 75, 133

ESD Effective Sampling Density. 16

fps frames per second. 35, 50, 100

FR Full Reference. 14, 15, 17

GUI Graphical User Interface. xix, 65, 181, 182

HDCA High Density Camera Array. xvi, xvii, xxii, 94, 97, 99–101, 106–108

HEVC High Efficiency Video Coding. xvi, xxi, xxii, 12–14, 16, 49, 50, 56, 58, 59, 83, 84, 95–97, 106, 114, 119, 123, 124, 129, 131, 152, 153, 155–157, 172, 174, 188

ITU International Telecommunication Union. 14, 26, 36–40, 43, 51–53, 55, 67–70, 86, 99–101, 118, 133, 134, 154, 175, 183, 188, 189

LAP Linear Approximation Prior. xxii, 153–157

LCD Liquid Crystal Display. 11, 77

LLE Locally Linear Embedding. 12, 84, 114, 124

MOS Mean Opinion Score. xv–xviii, xxii, 26–28, 30, 40, 42, 53, 55, 56, 59, 60, 62, 63, 69–71, 74–76, 86, 87, 90, 91, 101, 105, 108, 118, 121, 122, 134, 136–139, 163

MR-DIBR Multi Reference Depth Image-Based Rendering. 97, 98, 106, 166

MV-HEVC Multi View High Efficiency Video Coding. 13, 174

NR No Reference. 15

NTF Non-negative Tensor Factorization. 65

NU Nagoya University. 68–70, 74, 75, 133

OR Outlier Ratio. 40, 41, 70, 72–74, 134, 138, 139

PCC Pearson Correlation Coefficient. 40, 41, 64, 70, 72–76, 134, 138, 139

PSNR Peak Signal to Noise Ratio. xvi–xviii, xxii, 15, 16, 85–90, 98, 99, 102–107, 112, 117, 119–124, 132, 134, 135, 138, 139, 154–157

QoE Quality of Experience. 15, 17, 48, 65, 141, 164, 166

QP Quantization Parameter. xxi, xxii, 13, 49, 50, 96, 116, 152, 153, 174

QPI Quantum Photonic Imager. 11

RMSE Root Mean Square Error. 40, 41, 70, 72–74, 134, 138, 139

RR Reduced Reference. 14

SRCC Spearman's Rank Correlation Coefficient. 40, 41, 64, 70, 72–75, 134, 138, 139

SS Self Similarity. 12, 83, 84, 114, 124

SSIM Structural Similarity Index. xvi–xix, 15, 16, 86–90, 98, 99, 102–107, 132, 134, 135, 138, 139, 154–156

UHD Ultra High Definition. 94, 97

VALID Visual Quality Assessment of Light field Images Dataset. xix, xxii, 172, 173

VQM Video Quality Metrics. 189

VR Virtual Reality. 11

VVC Versatile Video Coding. 13

1 Introduction

Since its invention in the 19th century, photography has allowed to capture and share durable impressions of the world around us. The digitalization of photography and the widespread availability of cheap devices has made it a staple of everyday life, and a central part of our interaction with the world. According to 2016 Internet Trends report [Meeker, 2016], people upload on average 3.2 billion digital images per day. In addition, recent trends in image manipulation by consumers, such as Instagram filters and Snapchat lenses, show the growing need for new ways of interacting and engaging with the scenes to be captured.

The static nature of traditional photography poses several obstacles in nowadays scenarios. In particular, it limits the interaction with the captured scene, due to the capturing process itself. A digital camera records the intensity of light, focused by a lens or a series of lenses, hitting a photosensitive material (sensor); the 2D captured image retains no information about the position of objects in space or their depth. Additionally, the amount of light that is recorded by the sensor is heavily influenced by the lenses it has to pass through. This means that the aperture and the focal plane which were chosen at the moment of the capture cannot be modified after the acquisition.

Overcoming those limitations would impact two different market sections for cameras, albeit in different ways:

1. Professional photographers and operators may benefit from a tool that allows for greater flexibility when it comes to select the optimal parameters for shooting a scene. For example, an erroneous choice of focal plane in a scene may lead to several retakes and thus to greater expenses. Other features, such as change of point of view or zoom, may impact dramatically the way scenes are shot.

2. Consumers may look for an enhanced experience when capturing a special moment. Being able to change zoom, perspective and focus in a simple and intuitive way, without the need for expensive post-processing software. This is in line with the interactivity already seen in apps like Facebook or Snapchat, in which the users can modify the

appereance of the scene they have acquired with filters and lenses.

Light field photography proposes a new approach for the acquisition of scenes. The idea is to record all the information of light propagating through space, instead of just recording the amount of incident light in a particular point of the scene. This allows to render different views, with varying Depth Of Field (DOF) and focal planes, without the need of re-acquiring the scene.

Several techniques already exist to create light fields, and commercial solutions are available to consumers worldwide. However, the acquisition process creates more data when compared to traditional photography. This generates new challenges, especially regarding storage and trasmission of the light fields. Currently available handheld cameras, for example, use raw image formats, which are far from optimal when it comes to memory handling. New compression solutions that take into account the peculiarities of the data have been proposed; yet, no standard format has been adopted, which poses obstacles in interoperability, compatibility and spreading of light field photography. Moreover, most of the solutions that have been presented are hardly comparable with each other, due to differences in datasets and coding conditions that are used to assess their performance.

To reliably compare and evaluate the performance of compression algorithms for efficient encoding of light field data, adequate and reproducible subjective and objective quality evaluation methodologies are required. However, currently standardized recommendations for visual quality assessment, designed with traditional media in mind, are not equipped to cater to the complex nature of light field data. This is especially true in the case of subjective methodologies, which are designed for a passive fruition of the contents under test. Thus, by using the currently defined subjective methodologies, one may incur in the risk of neglecting the immersive and interactive nature of light field data.

In this work, we tackle the problem of visual quality evaluation, as well as compression, of light field contents. In particular, we first analyze current subjective quality assessment methodologies, and we design novel solutions to take into account the interactive nature of the data. We then perform benchmarking of several state-of-the-art strategies for light field compression, and we perform a thorough analysis on which approaches lead to the best coding efficiency. Lastly, we propose our own encoding scheme, showing improved performance with respect to the state of the art.

1.1 Contributions

Our contribution for this thesis can be classified in three main parts. The first part brings useful additions to the state of the art in light field quality evaluation, analysing legacy methodologies and proposing new procedures for subjective quality assessment of light field contents. The second part focuses on the comparison and evaluation of several coding solutions for

light field compression, highlighting the importance of reliable assessment in estimating the performance of different algorithms. Finally, the third part proposes a new encoding scheme to efficiently compress light field contents.

1.1.1 Visual quality assessment for light field contents

We present several methodologies for visual quality assessment of light field contents using image-based rendering. We first show that deploying single-image evaluation methodologies, which considerably increases the number of stimuli to be assessed, does not lead to a diversification in scores, and if not properly organised, can lead to biased ratings. Concluding that combining several renderings in one stimulus would be preferable, we then proceed to compare two methodologies for light field contents: one that allows interaction with the rendered images to favor engagement with the contents, and another that uses a prerecorded animation of the rendered views to ensure reproducibility and identical exposure of the contents to human subjects. Results show that, although the two approaches are strongly correlated, they are not statistically equivalent. In particular, the latter leads to smaller CIs and thus ensures more discriminative power among the tested solutions.

We then proceed to analyze user interaction patterns in light field visual quality assessment. We present several use cases in which analysis of user interaction can be beneficial and apport substantial advantages. We then perform a subjective quality test using five light field contents; during the test, user interaction is thoroughly tracked to provide data for the analysis. We show the results of the subjective tests, and we carry extensive analysis on the user interaction data. Moreover, we perform correlation between the two, to assess the predictive power of average interaction time for subjective visual quality scores. Results show that clear patterns can be seen in how users interact with the contents, and that strong correlation can be observed between the average interaction time for one content and its corresponding rating. In particular, we demonstrate that the average interaction time can be used to predict the subjective score of light field contents.

Finally, we present a thorough comparison of different quality assessment scenarios for subjective evaluation of light field contents on multi-layer tensor displays. We perform different sets of experiments in two separate laboratory settings, using both a prototype multi-layer display and a simulator for 2D screens. We propose two variants for double-stimulus subjective assessment, and we show that using one can be proven more beneficial when near-lossless levels of distortions are employed in the test. We also verify that poor correlation is achieved between sets of scores obtained with real multi-layer displays, when compared to a virtual simulator. Finally, we demonstrate that when the multi-layer display is used, the method employed to generate the layer patterns has a greater influence on the scores, with respect to the compression ratio.

1.1.2 Comparison and evaluation of compression solutions for light field contents

We first report the results of objective and subjective quality assessment performed under the framework of two grand challenges for compression of light field contents. We show that there is much to be gained in using new compression schemes as opposed to legacy JPEG and pseudo-temporal sequence-based video encoding. We also demonstrate that no proposed representation model is statistically better than the one adopted as reference.

We then present two different coding approaches for light field image compression, based on the information to be encoded. The approaches are evaluated using both objective quality metrics and ground truth scores from subjective experiments. The results provide some insights on the impact of compression algorithms and rate-reduction techniques, such as chroma subsampling, on the perceived visual quality of light field contents.

We also analyse different coding strategies for light field contents that will be rendered on multi-layer displays, through objective and subjective quality evaluations. For the latter, we employ both a software simulating multi-layer rendering on regular 2D monitors, and a prototype multi-layer display. We show that while one approach is significantly better than the others, results vary considerably depending on the rendering technology. In particular, we demonstrate that compression distortions are better perceived with the use of the simulator, whereas the prototype display is more sensitive to the method used to generate the layer patterns. Moreover, we verify that objective quality metrics are in alignment with the scores obtained with the prototype display, exhibiting a strong correlation. However, the correlation values between objective quality metrics and subjective quality scores drop when the simulator is used.

1.1.3 Towards new compression solutions for light field contents

We present a new approach to compress light field images based on a graph learning technique. While graph signal processing methods have been used in other works to improve the coding efficiency for light field images, the construction of the graph is usually imposed on the data. In our work, we focus on learning the graph in order to faithfully capture the similarities among perspective views. By using each view as a node, we considerably reduce the size of the graph while retaining its capability to capture variations among different views. We demonstrate its theoretical soundness, as well as its application to image coding. Our validating experiment shows that sensible gains can be achieved by using our solution against state-of-the-art encoders.

1.2 Organization of the thesis

The remainder of the thesis is organized as follows.

First, Chapter 2 explores preliminary concepts in plenoptic imaging and quality evaluation,

which represent the basis of our work. In addition, an overview of the state of the art in light field acquisition, rendering, compression and quality evaluation is given.

Part I focuses on the contributions to the field of subjective quality assessment for light field contents. In particular, in Chapter 3 we study single-image, interactive and passive methodologies for the subjective evaluation of light field contents, and we draw conclusions based on extensive statistical analysis. In Chapter 4, we investigate how user behavior can be influenced by the visual quality of the contents under assessment, and we show that a strong correlation can be observed between the time of interaction and the subjective ratings. Finally, in Chapter 5 we perform a comparison between different studies conducted on the visual quality of multi-layer displays, showing that additional parameters, such as the method for generating layer patterns, should be taken into consideration for rendering-dependent subjective experimentation.

Part II is dedicated to benchmarking of existing solutions for light field compression. More specifically, Chapter 6 presents the results of the evaluation campaigns conducted for two major Grand Challenges in light field compression, detailing how the assessment was carried out. Chapter 7 focuses on estimating the impact of using different coding approaches on the visual quality of light field contents, showing that traditional techniques for bitrate reduction, such as chroma subsampling, may lead to unwanted consequences on the perceptual quality of rendered contents. Lastly, in Chapter 8 we detail the results of applying different rendering-dependent compression strategies for light field contents visualized through a multi-layer display. We show that commonly used objective quality metrics, while sufficient for assessing the impact of the generation method for layer patterns on the visual quality, are not adequate for estimating the visual quality of light field contents under compression artifacts, at least when this type of rendering is involved.

Part III introduces new developments in light field compression. In particular, Chapter 9 presents a new compression solution for light field contents based on graph-learning techniques. We demonstrate that, by using the proposed method, remarkable improvement can be obtained with respect to state-of-the-art compression algorithms.

Chapter 10 summarises the contributions of the thesis, drawing some conclusions and recommendations from our work.

In the Annexes, selected contributions to open-source research can be found. In particular, Annex A describes a new dataset for quality assessment of light field images, providing both compressed and uncompressed contents, along with objective and subjective scores. Annex B presents a new framework for quality assessment of light field contents, which allows interaction with the contents while recording user behaviour. Annex C presents a software for performing quality assessment for field contents rendered through a tensor display.

2 Relevant work in light field imaging

2.1 Preliminary concepts

Although the idea of interpreting light as a field had been already proposed by Michael Faraday in a lecture entitled "Thoughts on Ray Vibration" in 1846, it was Andreī Gershun, in his book about radiometric properties of light in Three-dimensional (3D) space, who coined the term *light field* [Gershun, 1939]. Gershun defined the light field at each point as an infinite collection of vectors describing the amount of light that flows in every direction through every point in space.

The plenoptic function, first described in 1991 by Adelson and Bergen [Adelson and Bergen, 1991], defines a complete holographic representation of the visual world, describing the information available to an observer in any point in space and time. The function records the intensity of the light rays passing through every possible point in space (V_x, V_y, V_z) at every possible angle (θ, ϕ), for every wavelength λ at every time τ:

$$L = L(\theta, \phi, \lambda, \tau, V_x, V_y, V_z). \tag{2.1}$$

If one considers the light field at a fixed time $\bar{\tau}$, as for the acquisition of still images, then the plenoptic function can be represented as a 6D function. In the same way, considering the color sampling in *RGB* channels, one can discard the wavelength component λ. What is left is a function over a five-dimensional manifold of rays in free space:

$$L = L(\theta, \phi, V_x, V_y, V_z). \tag{2.2}$$

This is equivalent to the product of the 3D Euclidean space with a sphere. A representation of

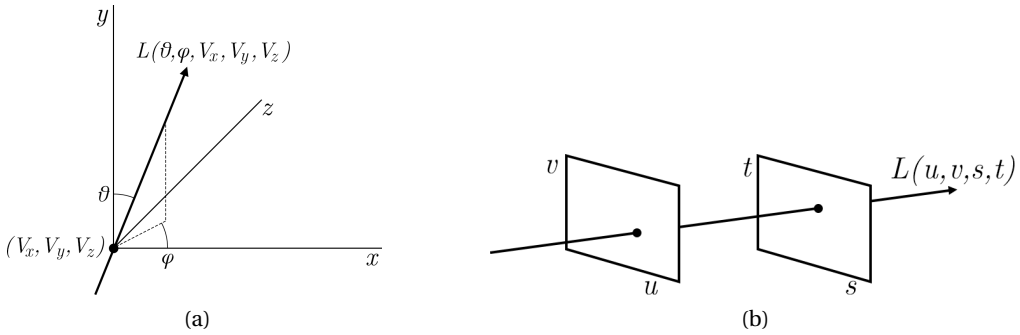


Figure 2.1 – Representation of plenoptic function in five dimensions (a) and four dimensions (b).

the 5D parametrization of the function can be seen in Figure 2.1 (a).

The 5D plenoptic function can be further simplified if one considers regions free of occluders (free space). This is due to the fact that radiance remains constant from point to point unless blocked. Exploiting the phenomenon, Levoy and Hanrahan proposed a new parametrization of the rays by their intersection with two planes in arbitrary position [Levoy and Hanrahan, 1996]. The coordinate systems of the two planes are (u, v) and (s, t), respectively. They restrict u, v, s and t to lie between 0 and 1, and they define an oriented line by connecting a point on the uv plane to a point on the st plane. The representation is called *light slab* (Figure 2.1 (b)):

$$L = L(u, v, s, t). \tag{2.3}$$

This parametrization allows to see the light field as a collection of perspective images of the st plane, each taken from an observer position on the uv plane. A light field can then be created from rendered images at fixed uv values, performing a sheared perspective projection so that each sample in the plane of the rendered image corresponds exactly to the st sample. This way, without any knowledge of the geometry of the scene, it is possible to render new views from the same scene, by extracting st slices from the appropriate uv point.

2.2 Light field acquisition

A digital 4D light field, which is a collection of perspective images, can be obtained by sampling the four-dimensional light field function defined in Equation 2.3. The density of the sampling in each dimension depends on the acquisition technology used to capture the light field image.

In general, different acquisition techniques can be used to capture light field images, depending on the requirements for baseline (i.e., the physical space that will be covered by the uv sampling), and for image resolution. More specifically, for a baseline in a range of meters, one

way of acquiring light field images is by means of a moving camera, which will acquire the different perspective views forming the 4D light field. In this case, the sampling in st plane depends on the camera resolution, and sampling in uv plane depends on the position of the capturing device and its shutter speed.
Examples of such acquisition devices are the Stanford Spherical Gantry, a motorized gantry with four degrees of freedom that can be used to capture light fields [Laboratory, 2004], and Apple's setup to construct 360-degree cylindrical panoramic images [Chen, 1995]. A robotized camera has been used to create the Fraunhofer Institute's light field dataset [Fraunhofer Institute, 2017]. Light fields can also be acquired by using hand-held cameras, as long as their position on the uv plane can be precisely estimated [Gortler et al., 1996][Buehler et al., 2001].

Another approach is to construct an array of cameras with synchronized shutter speed capturing the perspective views composing the light field all at once. In this case, the uv sampling depends on the baseline parameter of the camera array grid. Using a camera array, a full 4D light field is formed and new views corresponding to narrower baseline parameter must be further synthesized if needed.
An example of such acquisition technology is the Stanford Multi-Camera Array [Wilburn et al., 2005]. A distributed light field acquisition system, composed of 64 densely spaced video camera, is proposed in [Yang et al., 2002].

Light field images can also be acquired from multi-view plus depth data [Ouazan et al., 2011]. In this case, the baseline can be wide or narrow, depending on how the data was created [Zilly et al., 2012].

For light field image acquisition with narrow baseline, a hand-held plenoptic camera capturing so called "single lens stereo" can be achieved, by adding optical elements in front of the sensor plane in order to capture both angular and spatial information. In [Veeraraghavan et al., 2007], an attenuating mask is placed in the optical path of the main lens, exploiting heterodyning methods to map 4D rays into the 2D sensor. Liang et al. propose illumination multiplexing through a programmable aperture to capture light fields from a single camera, using sequential light field acquisiion [Liang et al., 2008]. Marwah et al. propose a new method to capture light field data from a single image, using optically-coded light field projections [Marwah et al., 2013]. Alam et al. propose a deconvolution approach to obtain narrow-aperture perspective views from a single wide-aperture image capture with a common camera [Alam and Gunturk, 2018].

A viable alternative to masking is by employing an array of micro lenses (lenslet) in front of the sensor plane. Two types of plenoptic cameras have been proposed in the literature. The first, called "unfocused light field camera" or "plenoptic camera 1.0", places the microlens array at the focal plane of the main lens [Adelson and Wang, 1992]. The spatial resolution (st plane) depends on the number of microlenses, whereas the angular resolution (uv plane) depends on the number of pixels behind each micro lens. The raw image obtained with this type of cameras closely resembles the honeycomb array of lenses that has been used for the

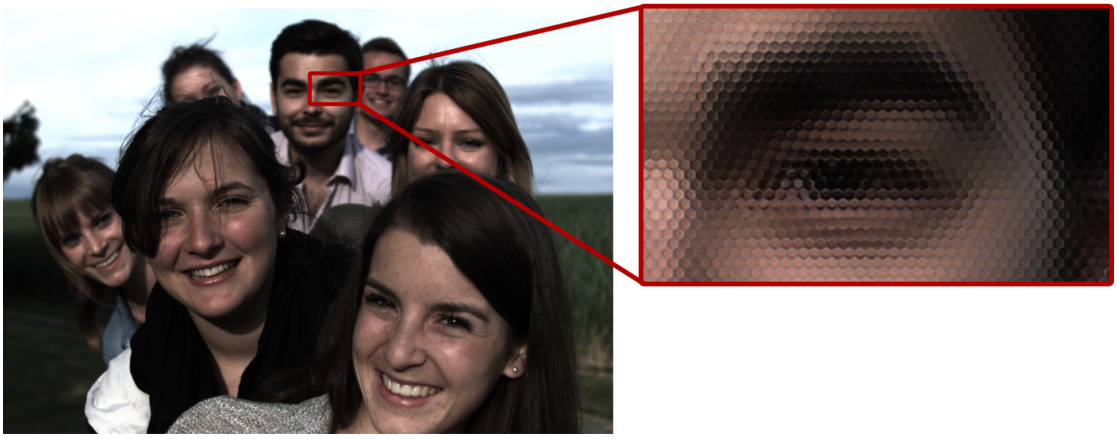


Figure 2.2 – Lenslet image obtained with an unfocused plenoptic camera. In the zoomed detail it is possible to observe the characteristic honeycomb structure, due to the microlens array.

acquisition, and will be from now on referred to as a lenslet image. Figure 2.2 depicts an example of lenslet image. It is possible to convert the lenslet image to a 4D data structure that effectively constitutes a sampling of Equation 2.3. The perspective views, in this case, can be obtained by selecting from each microlens the pixel supporting a certain viewpoint. Hand-held cameras implementing this model were presented in [Ng et al., 2005] and are already widely available to consumers[1]. Additionally, a light field microscope following the same principles has been proposed [Levoy et al., 2009, 2006].

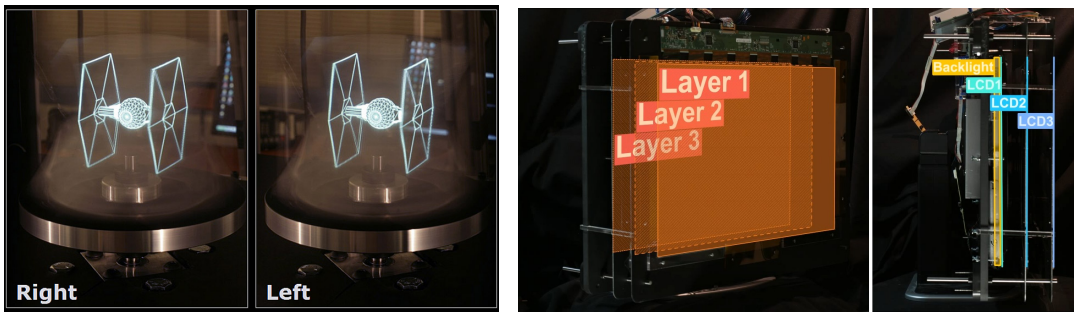
In the second type of plenoptic camera, namely, "focused light field camera" or "plenoptic camera 2.0", the sensor array is placed either before or after the microlens array's back focal plane [Lumsdaine and Georgiev, 2009]. This allows for a larger spatial resolution when compared to unfocused plenoptic cameras; however, the process of creating perspective views is not straightforward, as it requires to perform depth estimation [Palmieri et al., 2018]. Focused plenoptic cameras have been recently commercialized[2].

2.3 Light field rendering

The acquisition of any multimedia content would be fruitless without devices on which they can be enjoyed. In the case of light field photography, several displays have been proposed to overcome the limitations posed by traditional and stereoscopic monitors, by offering a more immersive experience of 3D contents with seamless transition among points of view and natural focal accomodation.

In the past, several multi-view displays have been presented that allow to visualize the scene from several points of view. Among other technologies, parallax barrier displays have been pro-

[1] https://www.lytro.com/
[2] https://www.raytrix.de/



(a) (b)

Figure 2.3 – Example of a 360° light field display (a) and of a multi-layer display (b). Courtesy of USC Institute for Creative Technologies, and MIT Media Lab, Camera Culture Group.

posed as a glass-free alternative to stereoscopic displays [Jacobs et al., 2003]. To allow for multi-view rendering, several technologies have been implemented, such as polarizer [Sakamoto and Morii, 2006] or Liquid Crystal Display (LCD) dynamic barriers through viewer tracking [Peterka et al., 2008]. Arrai et al. propose a 33-megapixel rendering system that uses a lens array to provide full parallax [Arai et al., 2010]. Balogh et al. use a system of projectors in combination with a holographic screen to fully render the 3D scene [Balogh, 2006]. Jones et al. offer a 360° light field rendering, using a highspeed video projector in combination with a spinning mirror [Jones et al., 2007] (Figure 2.3 (a)). El-Ghoroury and Alpaslan propose a new spatial light modulator, called Quantum Photonic Imager (QPI), for 3D rendering [El-Ghoroury and Alpaslan, 2014], and they demonstrate its application to light field displays [Alpaslan and El-Ghoroury, 2015].

A promising solution for light field rendering uses a stack of programmable light-attenuating layers in front of a light-emitting source, to provide depth cues without the need of glasses [Lanman et al., 2010] [Lanman et al., 2012] [Wetzstein et al., 2012]. The main advantage of this rendering method is that only a few attenuating layers are required to render multiple points of view; hence, the term “compressive display” has been used to define this type of devices. Figure 2.3 (b) depicts the structure of one multi-layer display with 3 layers. Multi-layer displays have been proposed for desktop applications [Kobayashi et al., 2017], and its technology has been adapted for big-screen light field projection [Hirsch et al., 2014].

More recently, a new generation of near-eye light field displays have been proposed, to experience full parallax multiview and focal accomodation in Virtual Reality (VR) and Augmented Reality (AR) scenarios. To do so, several technologies have been employed, including microlens arrays [Lanman and Luebke, 2013] [Wu et al., 2018], pinhole aperture arrays [Akşit et al., 2015], pupil tracking [Jang et al., 2017] and multi-layer attenuators [Maimone et al., 2013] [Huang et al., 2015].

Despite the rich literature on the topic, light field display technology is still not widely com-

mercialized. Available solutions include the Holovizio systems for back-projection displays[3], FoVI3D for 360°holographic rendering[4], and Avegant[5] and Magic Leap[6] for near-eye displays.

2.4 Light field compression

The increased capabilities that light field imaging provides come at the cost of the vast amount of data which is generated during the acquisition. Thus, several works have been focused on finding efficient compression algorithms to effectively store and transmit light field images.

The topic has attracted the attention of major standardization bodies. In 2014, the JPEG standardization committee launched a new initiative called JPEG Pleno, whose goal is to create a standard framework for efficient storage and delivery of plenoptic contents, including light fields, point clouds, and holograms. In particular, JPEG Pleno aims at finding the minimum number of representation models for these types of content, which, when necessary, can also offer interoperability with existing standards, such as legacy JPEG and JPEG 2000 formats. Since then, JPEG committee has been actively pursuing the definition of a new standard representation and compression algorithm for light field images [Ebrahimi et al., 2016].

Depending on the acquisition process, several approaches have been proposed. To allow random access while keeping the computational expenses low, Levoy and Hanrahan choose a two-stage pipeline, comprised of vector quantization and Lempel-Ziv coding [Levoy and Hanrahan, 1996]. Other works focuses on compressing synthetic 4D light fields using disparity compensation [Magnor and Girod, 2000, Jagmohan et al., 2003, Girod et al., 2003] and geometry estimation [Zhu et al., 2003].

More recently, the effort has been focused on compression of light field images acquired through hand-held devices. Several compression algorithms have been proposed to directly compress lenslet images through intra coding, exploiting redundancies in its structure (see Figure 2.2). For instance, Perra proposes a lossless compression scheme based on adaptive prediction [Perra, 2015]. Li et al. incorporates a full inter prediction scheme in HEVC intra prediction mode, explicitly exploiting the redundancy in lenslet images [Li et al., 2014b], as well as using the disparity compensation and impainting to efficiently code lenslet images [Li et al., 2016b]. Monteiro et al. introduce a modified version of HEVC Intra profile which integrates Locally Linear Embedding (LLE) and Self Similarity (SS) to improve block estimation [Monteiro et al., 2016]. More recently, Jin et al. propose a macropixel-based intra prediction method which first applies image reshaping to align the macropixel structure to the HEVC coding unit grid, and then defines three prediction modes to improve the coding efficiency of the blocks [Jin et al., 2018]. Schiopu et al. use a CNN-based approach to predict each macropixel from its neighbors and encode the residuals using context-based adaptive

[3] http://holografika.com/
[4] http://www.fovi3d.com/
[5] https://www.avegant.com/
[6] https://www.magicleap.com/

lossless coding [Schiopu and Munteanu, 2018].

Other solutions aim at improving the performance of existing video codecs by further exploiting the redundacies among the perspective views forming the 4D light field structure. A precursor of the approach is introduced by Olsson et al. [Olsson et al., 2006]. They propose the creation of sub-images from integral images. Such sub-images are then encoded through a pseudo-temporal sequence using AVC. Choudhury et al. adapt the method of coded snapshots to light field image compression through random codes [Choudhury et al., 2015]. Dai et al. code sub-aperture images using different scanning methods, including line and rotating scanning [Dai et al., 2015]. Helin et al. propose predictive coding for perspective views to achieve lossless compression [Helin et al., 2016, 2017]. Predictive coding is combined with segmentation-based context modelling by Schiopu et al. for lossless compression of lenslet images [Schiopu et al., 2017]. Ahmad et al. arrange the perspective views into a multiview structure that can be exploited by the corresponding extension of HEVC, namely MV-HEVC, and they propose a rate allocation scheme to optimize the performance by progressively assign the Quantization Parameter (QP) [Ahmad et al., 2017]. Jia et al. propose a fully reversible transformation to create the perspective views from the raw lenslet data [Jia et al., 2017]. The views are then optimally re-arranged and compressed using enhanced illumination compensation in an early version of the VVC software[7]. They also implement adaptive filtering to optimally reconstruct the lenslet image.

Several solutions have been proposed to exploit view synthesis or estimation to improve the coding efficiency. Jiang et al. use HEVC to encode a low-rank representation of the light field data, obtained by using homography-based low-rank approximation. They then reconstruct the entire light field by using weighting and homography parameters [Jiang et al., 2017]. Zhao et al. propose a novel compression scheme that encodes and transmits only part of the views using HEVC, while the non-encoded views are estimated as a linear combination of the already transmitted views [Zhao and Chen, 2017]. Astola et al. propose a method that combines warping at hierarchical levels with sparse prediction to reconstruct the 4D light field from a predefined set of perspective views [Astola and Tabus, 2018b,a]. The solution was recently adopted as the JPEG Pleno Verification Model [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018b]. Rizkallah et al. and Su et al. use CNN-based view synthesis to reconstruct the entire light field from 4 corner views, employing graph-based transforms [Rizkallah et al., 2018] or 4D-shape-adaptive Discrete Cosine Transform (DCT) [Su et al., 2018] to encode the residuals. Bakir et al. combine view estimation and view synthesis to reconstruct the 4D light field from a small set of reference views, which are encoded using the VVC software[8] [Bakir et al., 2018]. They use the linear approximation already seen in [Zhao and Chen, 2017] to reconstruct a subset of the views, and then employ a CNN-based approach to reconstruct the remaining perspective views.

Other approaches exploit the redundancy of the light field data through novel representation.

[7] https://jvet.hhi.fraunhofer.de/
[8] https://jvet.hhi.fraunhofer.de/

Verhack et al. use a multi-modal Gaussian Mixture Model to represent the light field data, relying on perceptual similarity more than exact reconstruction [Verhack et al., 2017]. The perspective views composing the light field data can be reconstructed at the decoder side using only the parameters of the model. Su et al. propose a graph-based representation to perform multi-view prediction from a single image, using HEVC to encode the residuals [Su et al., 2017]. De Carvalho et at. propose the adoption of 4D DCT to obtain a compact representation of the light field structure [de Carvalho et al., 2018]. The DCT coefficients are grouped using hexadeca-trees, for each bitplane, and encoded using an arithmetic encoder. Komatsu et al. present a novel coding scheme based on weighted binary images [Komatsu et al., 2018]. The 4D light field structure is reconstructed from a set of binary images, which are predefined for all the perspective views, along with a set of weights, which may vary from view to view.

2.5 Light field visual quality evaluation

Evaluation of visual quality as perceived by the end users is of paramount importance in determining the efficacy of the proposed solutions for rendering, compression and processing of light field contents. In terms of the consequences of compression algorithms on visual quality, most solutions listed in Section 2.4 provide a preliminary performance evaluation. However, the evaluation procedures, not to mention the coding conditions, are usually divergent among different publications. Thus, drawing a straightforward comparison between distinct solutions can be proven challenging.

Generally, visual quality assessment can be carried out with either subjective evaluations or objective quality metrics. In the first case, users are directly probed regarding the level of perceived degradation or goodness of the content under assessment. Guidelines have been drafted by standardization committees to ensure that the information gathered with subjective tests would produce statistically relevant results, without unwanted bias. In particular, the International Telecommunication Union (ITU) agency provides recommendations regarding subjective assessment of visual quality for television pictures [ITU-R BT.500-13, 2012], multimedia applications [ITU-T P.910, 2008], and video, audio and audiovisual quality of internet video and distribution quality television in any environment [ITU-T P.913, 2016]. As a comprehensive overview of all recommendations for subjective evaluation would be out of the scope of this thesis, we refer interested readers to [Perrin, 2019].

Subjective evaluation, while providing ground truth information regarding the perceived visual quality of the contents under assessment, is often burdensome and costly. Thus, objective quality metrics have been designed to estimate the level of impairment of a given stimulus. Objective quality metrics are mathematical models approximating perceptual responses. They are commonly classified into three categories, depending on the degree of reference information on which they depend on: Full Reference (FR) metrics, which perform an estimation of the visual quality of an impaired content with respect to its source reference; Reduced Reference (RR) metrics, which take advantage of some features of the source reference to

perform an assessment of the visual quality of the impaired content; and No Reference (NR) metrics, which determine a visual quality score based on the impaired content, without any reference information. Among the three alternatives, FR metrics are the most commonly used to assess the visual quality of image contents under compression distortions. In particular, Peak Signal to Noise Ratio (PSNR), a widely used metric to measure the level of noise on a pixel color basis, has been consistently employed to report on the performance of compression solutions for image and video contents, and was promptly adopted to describe the coding efficiency of novel methods for light field compression (see [Levoy and Hanrahan, 1996]). The JPEG Pleno Common Test Conditions [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018a], defined in July 2018 by the JPEG standardization body, adopt PSNR as performance metric for the core experiments, along with Structural Similarity Index (SSIM) [Wang et al., 2004], a metric modeled on human visual perception of luminance, contrast and structure.

Subjective quality assessment of light field contents on light field displays has been prominently featured in several publications. Spatial resolution of back-projected light field displays has been investigated by Kovacs et al. [Kovács et al., 2014]. In particular, the authors examine how viewing angle affects the perception of spatial resolution, along with the role played by motion parallax. Darukumalli et al. inspect the relationship between zooming levels, region of interest and subjective quality of light field contents, using Absolute Category Rating (ACR) and Degradation Quality Rating (DQR) [Darukumalli et al., 2016]. Kara et al. analyse the correlation between spatial and angular resolution, and how reducing spatial resolution can improve parallax perception [Kara et al., 2017a]. They also study the impact of angular resolution on the perception of light field content, first in a free movement scenario, and then with fixed observer position, using ACR [Kara et al., 2016, 2017b]. Adhikarla et al. perform subjective evaluation on a 3D monitor with head tracking to simulate parallax effect, using pairwise comparison, to assess the performance of various objective quality metrics on distorted light field contents [Adhikarla et al., 2017]. Selected distortions include compression, interpolation, warping and rendering artifacts. Similarly, Shi et al. analyse the performance of objective quality metrics by comparing them with subjective scores obtained on a 3D monitor setup, using a newly-defined windowed 5-degree-of-freedom light field image database [Shi et al., 2018]. Tamboli et al. investigate the Quality of Experience (QoE) associated with back-projected light field displays, using ACR [Tamboli et al., 2018a].

As light field displays are scarcely available, image-based rendering has been used to perform subjective assessment on 2D screens. Paudyal et al. investigate the impact of watermarking on visual quality of light field contents, and especially the relationship between watermark strength and visual quality, using ACR [Paudyal et al., 2015]. Filipe et al. [Filipe et al., 2018] assessed the performance of several state-of-the-art focus metrics in the evaluation of extended DOF (all-in-focus) images acquired by a focused plenoptic camera. In particular, they performed a subjective assessment in which they compared extended DOF images obtained using optimal patch sizes, against images obtained with slightly larger patch sizes. Paudyal et al. [Paudyal et al., 2017a] analysed the impact of different visualization techniques, including image-based assessment of all-in-focus perspective views and refocused views, using

ACR, concluding that there is high correlation between the scores obtained by image-based evaluation when compared to the corresponding animation-based passive evaluation. Perra et al. analyse how light field subsampling affects the perceived quality of refocused views, which are presented in an animated fashion [Perra et al., 2018]. Battisti et al. compare several visualization techniques, as well as different framerates, for light field quality assessment, concluding that horizontal scan is to be preferred [Battisti et al., 2018]. Upenik et al. adapt the image-based representation for omnidirectional visualization, using a head-mounted display [Upenik et al., 2018].

A few publications have been devoted to assess the visual quality of light field contents using conventional objective quality metrics. Vieira et al. compare five different HEVC-compatible coding of lenslet images with different data formats, using PSNR-YUV [Vieira et al., 2015]. Rizkallah et al. report the impact of compression of light field images on refocusing and extended focus images through objective quality metrics such as PSNR and SSIM, and they propose a new metric to measure the amount of compression blur in focused regions [Rizkallah et al., 2016]. Perra et al. analyze the effect of HEVC-based compression on light field refocusing, using metrics and conditions defined in [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018a] to report the tradeoff between compression and objective quality [Perra and Giusto, 2018]. Alves et al. detail the results of a performance assessment campaign for light field image compression on several rendered images, using PSNR-Y as metric [Alves et al., 2016].

New objective quality metrics and evaluation frameworks have been proposed to better model the peculiarities of light field imaging. Niemann et al. perform an evaluation on computing light fields from a hand-held camera, without any further input [Niemann and Scholz, 2005]. They evaluate the accuracy of estimating parameters for light field using feature selection and tracking, factorization of one initial sequence and depth map estimation. They also introduce a new method for measuring the quality of a light field. Ramanathan et al. create a framework for analysing the impact of various parameters on the rate-distorsion curve of light field coding [Ramanathan and Girod, 2006]. The parameters include correlation within an image and between images, geometry accuracy and prediction dependency structure. Shidanshidi et al. also introduce a new objective quality metric to perform an evaluation of light field rendering [Shidanshidi et al., 2011a] [Shidanshidi et al., 2011b]. They present a geometric measurement, called Effective Sampling Density (ESD), and they perform a comparison on existing techniques for interpolation of new views from light fields. Fu et al. analyse the effects of light field photography in image quality [Fu et al., 2011]. They first develop a simulation approach to test visual resolution and other image quality evaluation metrics for light field photography. Then, they compare the results with conventional cameras to discuss improvements and shortcomings of light field photography. Meng et al. evaluate the performance of a multi-spectral plenoptic prototype camera by introducing new performance metrics [Meng et al., 2013]. The metrics aim at assessing the spectral reconstruction quality. Jarabo et al. uses a different approach, and evaluates different light field editing interfaces, tools and workflows [Jarabo et al., 2014]. The experiment aims at evaluating the difficulty of performing modification on light field images from a user perspective, and how reconstructed

depth maps influence the task. Tamboli et al. propose a new 3D FR metric that combines a spatial component and an angular component to measure the objective visual quality of 3D contents on light field displays [Tamboli et al., 2016]. The metric is used in [Tamboli et al., 2018b] to evaluate the quality of key-frames from light field video contents.

2.6 Summary and perspectives

In this section we have provided a concise overview of the ongoing work in rendering, compression and quality evaluation for light field contents, after introducing some preliminary concepts on acquisition and representation. Although the amount of work devoted to the topic is remarkable, we believe there is still a number of issues that need to be addressed and fixed. Our work in this thesis aims at covering some of the inadequacies, listed as follows:

1. Image-based rendering is a cost-effective solution to experience light field contents on 2D screens. Although both objective [Alves et al., 2016] and subjective [Paudyal et al., 2017a] evaluations have been conducted in the past using selected rendered contents from the light field data, no extensive study has been carried on the difference between the perceptual quality of single rendered images on 2D displays. Moreover, the near totality of the work on subjective evaluation using image-based rendering is displaying light field contents as traditional media items, such as still images or prerecorded videos. Interactivity with the content is almost completely disregarded ([Shi et al., 2018] being a notable exception). We aim at bridging the gap between single-image and multi-view assessment in Chapter 3, and we provide extensive analysis of interaction patterns in Chapter 4.

2. While extensive work has been conducted on subjective assessment of back-projected light field displays, the same cannot be said for other types of light field displays. Due to their limited availability and their hardware limitations, multi-layer displays have not been employed in subjective evaluation campaigns, and the QoE associated with them is still largely an open problem. We propose a framework to simulate multi-layer rendering on traditional 2D screens (Annex C), and we perform a comparison among different test conditions for light field interactive assessment in Chapter 5. In Chapter 8 we assess the performance of compression solutions for multi-layer displays through both subjective and objective means.

3. The selection of different coding conditions for performance evaluation of compression solutions for light field contents makes it hard, if not impossible, to perform a comparison among them. The adoption of common test conditions, such as the ones detailed in [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018a], is an auspicable step in achieving clear and straightforward benchmarking of distinct algorithms. We believe reporting the evaluation results of two major grand challenges on light field compression, as we do in Chapter 6, will help identifying the best performing approaches to foster research

in the field.

4. Although video encoding was promptly adopted as a viable solution to efficiently compress light field data, as seen for example in [Olsson et al., 2006] and [Li et al., 2014b], little work has been conducted in assessing whether common stream reduction techniques, such as chroma subsampling, can be transmuted to light field coding. We address the issue in Chapter 7, where we also compare intra and inter approaches for lenslet image compression.

5. View estimation and synthesis seem to be among the most promising solutions for light field compression. We contribute to the ongoing effort in finding the best representation for disparity data in Chapter 9, where we propose a lightweight estimation of view interdependency using graph learning techniques.

Part I

Visual quality assessment for light field contents

3 Analysis of different methodologies for image-based light field quality assessment

Disclaimer: Some of the contents of this chapter were adapted from the following articles, with permission from all co-authors and publishing entities:

Viola, Irene, Martin Řeřábek, Tim Bruylants, Peter Schelkens, Fernando Pereira, and Touradj Ebrahimi. "Objective and subjective evaluation of light field image compression algorithms." In Picture Coding Symposium (PCS), 2016, pp. 1-5. ©2016 IEEE.

Viola, Irene, Martin Řeřábek, and Touradj Ebrahimi. "Impact of interactivity on the assessment of quality of experience for light field content." In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pp. 1-6. ©2017 IEEE.

Personal contribution: The subjective quality assessment tests were designed with the help of my co-authors. I performed the experiments and curated the analysis.

For any type of multimedia content, reliable quality assessment is of paramount importance in the design and validation of new compression solutions that aim at reducing the size of the original data without compromising its perceptual quality. While objective quality metrics have been developed in the last decades to effectively predict the perceptual quality of the contents under assessment, subjective quality evaluations remain the most reliable means to measure the quality of media contents. However, quality assessment of light field contents poses new questions and challenges, due to the enriched nature of the content and the possibilities it offers for the rendering step.

One of the most natural and intuitive ways to consume light field contents would involve light field displays or simulators to create a multi-view, 3D rendering of the contents [Balogh, 2006, Matsubara et al., 2015]. Using this approach, the full potential of light field imaging is exploited to create a 3D representation of the scene in front of the user. However, such displays are not widely available to consumers, due to their cost and their requirements.

Another possible approach to render light field contents relies on image-based rendering to showcase the increased capabilities of light field contents [McMillan and Bishop, 1995].

With this approach, the light field information can be sampled and combined to create 2D images, which can be displayed on regular displays. For example, a perspective image can be created by selecting a specific point in the (s, t) plane, as defined in Equation 2.3. It is also possible to combine different perspective views to change the focal plane in the scene (digital refocusing). The range of possibilities is virtually endless, which poses the problem of what rendered images should be used when performing visual quality assessment, as well as how to properly present them to the user in a subjective quality evaluation scenario.

The simplest way to perform an evaluation of light field contents using image-based rendering is by evaluating each rendered image separately, using common image visual quality evaluation techniques. A final score for the entire light field content under examination can be obtained by computing the mean of all the scores obtained by the single rendered images, for example. Although fairly simple and straightforward to implement, this evaluation methodology creates an overwhelmingly large amount of stimuli to be processed and assessed, thus adding to the strain of already costly evaluation campaigns. On the other hand, evaluating each rendered image individually allows to account for possible quality variations among different renderings, which could be proven useful for both encoding and rendering purposes.

The most natural way for the user to exploit these possibilities is by interacting with the content. Indeed, being able to change the appearance of the scene that has been acquired is a desirable feature, one that is already implemented in widespread applications such as Instagram or Facebook. From this perspective, interactive methodologies for subjective quality assessment should be actively deployed since they give a more accurate depiction of how the user consumes and engages with the content. However, one significant shortcoming of the interactive approach is the lack of control on what users are visualising and thus what is being rated. Since each subject decides autonomously which rendered image to display and for how long, there is little control over the number of rendered images that each subject is examining, nor there is guarantee that different subjects are visualizing the same set of images.

An alternative way to evaluate visual quality of light field contents would be to use a passive approach, where the subjects are presented with a pre-recorded animation displaying different rendered images. Such an approach guarantees that each subject sees the identical set of images, rendered with the same parameters, under the same conditions. However, to yield reliable results, a number of parameters should be carefully selected, such as the optimal framerate and the number of rendered images to be presented to the subject. Moreover, a passive approach disregards the interactive nature of light field contents, and thus does not always faithfully represent the average user experience in consuming light field contents.

In this chapter, three quality evaluation methodologies specifically designed for light field using image-based rendering will be presented. The first one evaluates the rendered images obtained from each light field content separately. We analyse the results of the subjective quality evaluation campaign by means of statistical tools, to understand whether the advantages of evaluating the rendered images separately compensate the additional strain. We then

move on to compare the second evaluation methodology, which enforces interaction with the content, with the third methodology , which employs a prerecorded animation of rendered images, to ensure the same experience for all users. We finally draw conclusions from our analysis and provide recommendations for future tests.

3.1 Single image evaluation for light field contents

Image-based rendering offers an impressive showcase of the rendering abilities made possible by light field technology. Among other possibilities, the point of view of the scene can be modified, digital refocusing can be applied to highlight a specific plane in the image, zooming can be performed to exclude some planes in the scene, and so on. Due to the nearly endless possibilities that are offered, the first challenge that is poised for any type of image-based light field evaluation is to select which rendering to take into account when evaluating the contents. Such challenge becomes particularly dire when the images obtained after the rendering procedure are evaluated singularly, as the length and complexity of the test grows accordingly. Equally challenging is obtaining a single score for each content under assessment, since multiple rendering are presented and subsequently rated in a separate fashion.

Image-based assessment has been used in literature to subjectively evaluate the quality of light field contents. Filipe et al. [Filipe et al., 2018] assessed the performance of several state-of-the-art focus metrics in the evaluation of extended DOF (all-in-focus) images acquired by a focused plenoptic camera. In particular, they performed a subjective quality assessment in which they compared extended DOF images obtained using optimal patch sizes, against images obtained with slightly larger patch sizes. Paudyal et al. [Paudyal et al., 2017a] analysed the impact of different visualization techniques, including image-based assessment of all-in-focus perspective views and refocused views, using ACR, concluding that there is high correlation between the scores obtained by image-based evaluation when compared to the corresponding animation-based passive evaluation. Image-based assessment was chosen as the subjective quality evaluation methodology for the ICME 2016 Grand Challenge on Light Field Coding [Viola et al., 2016a]. Five algorithms were received as response to the challenge, and were evaluated against the anchor of choice, namely JPEG, using both objective and subjective quality assessment. For the latter, six light field contents were chosen among the dataset. For each content, five rendered images were created, comprising three all-in-focus perspective views and two refocused views, to carry the subjective test. The number of possible renderings was purposedly constrained to avoid an overly complex assessment scenario; even so, a total number of 720 stimuli was generated for the evaluation campaign.

Hence, it is crucial to analyse the results of such a massive campaign, to assess whether the added complexity of the test was justified by a diversification in ratings among different rendering procedures. To this aim, in this chapter we use statistical tools such as Analysis of variance (ANOVA) to examine the similarities among the ratings, and in particular to determine whether there are statistically relevant differences among different rendered images. Results

are decisive in selecting the best evaluation methodology for visual quality assessment of light field contents.

This chapter focuses on analysing shortcomings and advantages of using image-based assessment to evaluate light field contents, using as benchmark the results of the ICME 2016 Grand Challenge campaign. In the following subsections, we will describe the subjective quality evaluation methodology in details, as well as the statistical tools used to perform the analysis, and we will present the results of our inquiry. For a thorough presentation of the grand challenge results, including comparisons among the codecs and guidelines on different approaches, we refer interested readers to Chapter 6.

3.1.1 Data preparation and coding conditions

As input for the grand challenge, light field images created with a Lytro Illum plenoptic camera were selected. In particular, proponents were asked to compress a lenslet image, which was created from the raw 10-bit sensor data by applying devignetting, demosaicing, clipping to 8-bit and color space conversion from RGB444 to YUV420.

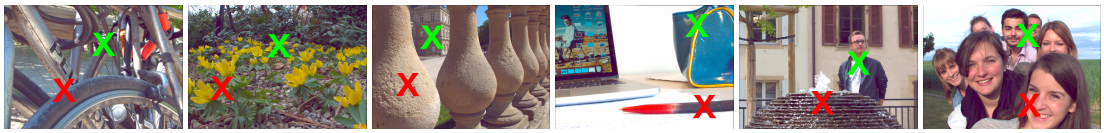
The performance of the proposed compression algorithms was evaluated at four fixed compression ratios, namely $R1 = 10:1$ (1 bit per pixel (bpp)), $R2 = 20:1$ (0.5 bpp), $R3 = 40:1$ (0.25 bpp), $R4 = 100:1$ (0.1 bpp). The ratios were computed with respect to the size of the raw data obtained from the camera.

For the objective and subjective quality evaluations, the decompressed lenslet image was converted to a stack of all-in-focus perspective views (light field data structure) using the Matlab implementation of the Light Field Toolbox v0.4 [Dansereau et al., 2013][Dansereau et al., 2015]. Each perspective view was created by selecting and aligning samples from each micro-lens element that supported a particular point of view. The resulting light field data structure is a 5-D array with dimensions of $15 \times 15 \times 434 \times 625 \times 3$, in which 15×15 is the number of perspective views, 434×625 is the resolution of each view, and 3 corresponds to the color channels. Color and gamma corrections were applied to each perspective view.

The same pipeline was employed to generate the reference light field data structure from the uncompressed YUV420 lenslet image.

A total of six light field images were selected from a Lytro Illum lenslet database [Řeřábek and Ebrahimi, 2016] for the subjective quality assessment. The central view of each content is depicted in Figure 3.1.

For each content, three all-in-focus perspective views were directly extracted from the light field data structure. In particular, from the 15×15 stack of perspective views, the ones at indexes $(8, i)$, where $i = 5, 8, 11$, were selected to represent different perspectives of each scene. Additionally, the MATLAB toolbox was used to create two refocused views for each light field



(a) Bikes (b) Flowers (c) Stone_Pillars_Outside (d) Desktop (e) Fountain_&_Vincent_2 (f) Friends_1

Figure 3.1 – Central all-in-focus perspective view from each content used in the experiments. Refocused points marked in green (slope 1) and red (slope 2). ©2016 IEEE

Image ID	Slope 1	Slope 2
Bikes	-0.65	0.22
Flowers	-0.3	0.3
Stone_Pillars_Outside	-0.5	0.2
Desktop	-0.5	0.5
Fountain_&_Vincent_2	-0.5	0.35
Friends_1	-0.15	0.2

Table 3.1 – Values of slope for refocused views. ©2016 IEEE

content, using a modified version of the function *LFFiltShiftSum*. The function shifts all the perspective views according to a parameter, referred to as a slope, which determines the focal plane. A sum of the shifted views is performed in order to obtain a single image that is refocused on a specific plane, depending on the value of the slope. The number of views to be shifted and consequently summed defines the DOF. Summing all 15×15 views creates the smallest DOF, in which only one specific plane in the image is in focus. On the other side, taking just the central perspective view, which is equivalent to summing just 1×1 views, brings all the objects in focus (largest DOF). For the test, it has been chosen to sum perspective views from index 5 to index 11 (7×7 views) in order to have a larger DOF that still showed the effects of refocusing. Two slopes were selected in order to focus the image on two different planes in the scene. Figure 3.1 illustrates the chosen points for refocusing (Slope 1 in green, Slope 2 in red). The values of the slope parameter used in the function are listed in Table 3.1. The three all-in-focus perspective views, along with the two refocused views, form five views per content.

In total, six algorithms (five proponents and one anchor) were evaluated in the subjective quality assessment tests. Five rendered views were created from six light field contents compressed at 4 different bitrates. Thus, a total of 720 stimuli was evaluated in the test.

Subjective quality assessment methodology

The methodology selected to conduct the subjective quality tests is based on Double Stimulus Continuous Quality Scale (DSCQS). Two images in native resolution (625×434 pixels) were presented simultaneously in a side-by-side fashion. One of the two images was always the

Table 3.2 – Test environments and specifications.

Approach	Environment	No. contents	No. codecs	No. bitrates	No. subjects	Methodology	No. persp. views	No. refoc. views
Single-stimulus	Semi-controlled crowdsourcing	6	6	4	35	DSCQS	3	2

uncompressed reference, and its position on the screen was randomized. The other image was compressed by one of the evaluated algorithms at one of the evaluated bitrates. The same rendering parameters were used for both the reference and the test images. Subjects were asked to rate the quality of both images on a discrete scale from 5 (Excellent) to 1 (Bad). They were informed that one of the images was the reference, but they did not receive any indication on its relative position on the screen. Before the experiments, a training session was organized to help subjects to adjust to the peculiarities of light field rendering, and to help them to detect various distortions and compression artifacts. Five training samples were generated using an additional content from the light field database [Řeřábek and Ebrahimi, 2016]. To perform the tests, the QualityCrowd 2 framework [Keimel et al., 2012] was modified to suit the DSCQS methodology.

The experiment was split into four sessions. In each session, 180 pairs of images were shown, corresponding to approximately 45 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in two sessions. A break of ten minutes was enforced between the sessions to avoid fatigue. At the beginning of first session, one dummy example was shown to ease the subject into the task. The resulting scores from dummy stimuli were not included in the final results.

Overall, 35 naïve subjects (24 males and 11 females) participated in the subjective experiments, each rating 360 stimuli over the course of two sessions. Subjects were between 18 and 33 years old. The average and median age were 22.4 and 22 years old, respectively. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts. A summary of the specifications of the test can be found in Table 3.2.

3.1.2 Data processing and statistical analysis

Outlier detection and removal was performed according to the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. One subject was found to be an outlier, and the corresponding scores were discarded. This lead us to 17 scores per stimulus. After outlier removal, the Mean Opinion Score (MOS) was computed for each coding condition j (i.e. for each content, view, proponent and bitrate) as follows:

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} m_{ij} \tag{3.1}$$

where N is the number of subjects and m_{ij} is the score for stimulus j by subject i. The corresponding 95% Confidence Intervals (CIs) were computed using Student's t-distribution.

Analogously, for each reference stimulus a MOS score was computed:

$$MOS_{\bar{j}} = \frac{1}{N}\sum_{i=1}^{N} m_{i\bar{j}} \tag{3.2}$$

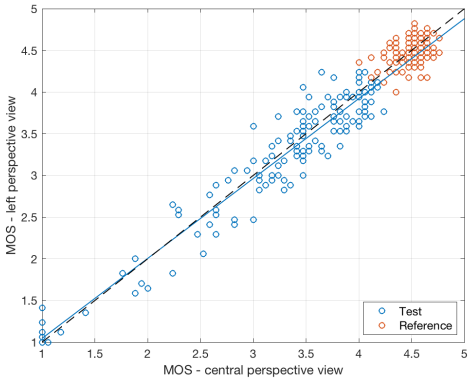
in which $\bar{j}$ represents each reference condition (i.e. each content and rendered view).

In order to determine whether statistically significant differences are present among the ratings given for differently rendered images, we performed a multiway ANOVA on the data. In particular, we first performed a one-way ANOVA on the MOS scores associated with each coding condition j, analysing whether there was any statistical difference associated with different rendering parameters. We then performed the same analysis on the MOS scores associated with each rendering condition $\bar{j}$, in order to further understand whether any difference in performance is due to coding artifacts. Finally, we performed an n-way ANOVA on the full set of scores to gain insights on the statistical differences within the coding conditions.

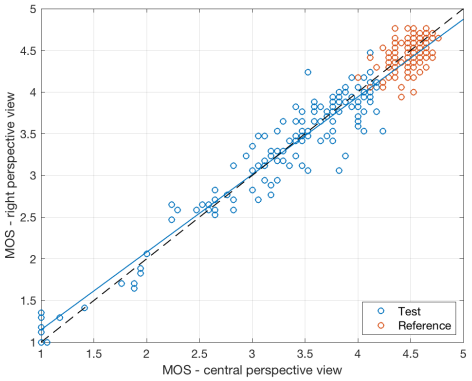
3.1.3 Results and discussion

Figure 3.2 compares the MOS scores given to each perspective and refocused view, for each test content and respective reference. As further showed by the linear fitting for test contents, the test scores are evenly distributed along the $y = x$ line, proving that strong correlation can be found between the scores assigned to perspective and refocused views, within their group. To further demonstrate that different perspective and refocused views were scored similarly within their group, we perform a one-way ANOVA on the test scores, using as discriminative value the corresponding perspective (central, left or right) or refocused (front or back) view. Figure 3.3 and Tables 3.3 and 3.4 show the results of the analysis. As made clear by the high p-values (0.3071 and 0.341 for the perspective and refocused views, respectively), the scores assigned to test contents rendered through different perspective and refocused views were statistically equivalent at 1% significance level. Thus, only one representative of each group could have been used in the test, sensibly reducing the complexity and length of the evaluation, without causing any distruption in the collected scores. Results from one-way ANOVA applied on the reference data show similar trends, reporting p-values well above the significance threshold ($p = 0.0166$ and $p = 0.772$ for perspective and refocused views, respectively).

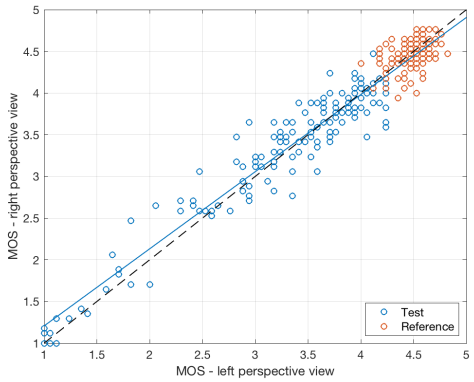
Once the correlation within the groups of views has been analysed, we investigate whether there is any difference to be found between the two groups. Figure 3.4 shows the comparison



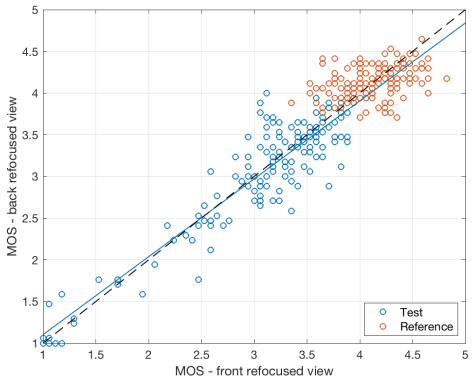
(a) Central vs left perspective view



(b) Central vs right perspective view



(c) Right vs left perspective view

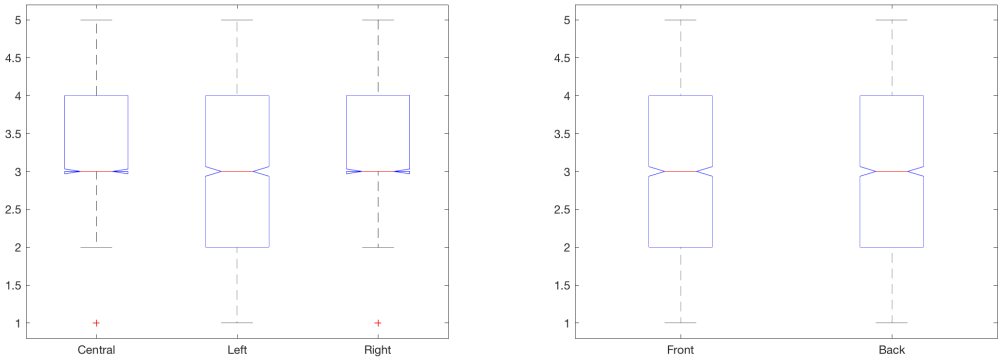


(d) Front vs back refocused view

Figure 3.2 – Comparison between MOS values for different perspective views and different refocused views, for test contents (blue) and respective references (orange), with relative linear fitting. The dashed black line represents the $y = x$ function.

between the MOS values assigned to all the perspective views, with respect to the MOS values associated with the refocused views. As showed in the plot, the vast majority of points fall below the $y = x$ line, signifying that the scores assigned to the perspective views were steadily higher than their refocused counterpart. Interestingly enough, the same trend can be observed not only for the test contents, but for the references as well. Despite being trained on considering only the differences between test and reference images, subjects consistantly gave lower ratings to both test and reference contents when presented with refocused views, whereas for perspective views the ratings were usually higher. Results from the one-way ANOVA, summarised in Tables 3.5 and 3.6, confirm that the two groups have statistically significant differences ($p = 1.78763\text{e}{-25}$ and $p = 3.06596\text{e}{-121}$ for test and reference contents, respectively).

In order to assess the relevance of each coding condition on the set of scores, we perform multi-



(a) Perspective views

(b) Refocused view

Figure 3.3 – Boxplot analysis of the raw scores assigned to each perspective and refocused view.

Table 3.3 – One-way ANOVA on the raw scores given to test contents rendered through perspective views.

	SumSq	**DF**	**MeanSq**	**F**	**p-value**
Perspective views	3.3	2	1.63249	1.18	0.3071
Error	10148.4	7341	1.38243		

Table 3.4 – One-way ANOVA on the raw scores given to test contents rendered through refocused views.

	SumSq	**DF**	**MeanSq**	**F**	**p-value**
Refocused views	1.15	1	1.1489	0.91	0.341
Error	6199.85	4894	1.26683		

way ANOVA on the test and reference scores, considering two-factor interactions. We first consider the scores assigned to perspective and refocused views separately. Tables 3.7, 3.8, 3.9 and 3.10 show the results of the analysis. As seen before, the groups fall above the 1% significance threshold ($p = 0.1081$ and $p = 0.0156$ for scores assigned to rendered perspective views in test and reference contents, respectively, whereas for reference views the results are $p = 0.2273$ and $p = 0.7686$ for test and reference contents, respectively). Among the first order interactions, it is worth mentioning that the interaction between contents and refocused views is significant for both test and reference contents, meaning that particular combinations of the two influenced how the stimuli were scored.

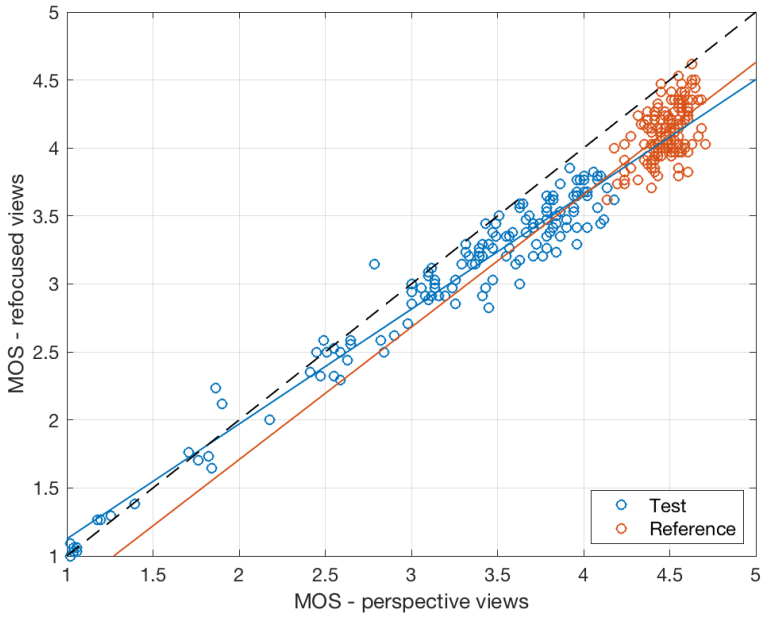


Figure 3.4 – MOS values for perspective views vs MOS values for refocused views, for test contents (blue) and respective references (orange), with relative linear fitting. The dashed black line represents the $y = x$ function.

Table 3.5 – One-way ANOVA on the raw scores given to test contents, divided by type of view.

	SumSq	**DF**	**MeanSq**	**F**	**p-value**
Perspective and refocused views	146	1	146.046	109.3	1.78763e−25
Error	16352.7	12238	1.336		

Table 3.6 – One-way ANOVA on the raw scores given to reference contents, divided by type of view.

	SumSq	**DF**	**MeanSq**	**F**	**p-value**
Perspective and refocused views	373.88	1	373.878	560.71	3.06596e−121
Error	8160.21	12238	0.667		

Finally, we perform multi-way ANOVA on the entire set of scores, considering both test and reference contents simultaneously. Table 3.11 summarises our findings. As can be seen, the scores assigned to different views are to be considered significantly different in a statistical sense. Moreover, the interaction between contents and views and codecs and views is statistically significant. The latter is especially important, because it signals that the choice of rendering parameters can influence how different codecs are assessed, indipendently of the

Table 3.7 – Multi-way ANOVA (interaction model) on the raw scores given to test contents (perspective views only).

	SumSq	DF	MeanSq	F	p-value
Contents	658.3	5	131.651	179.48	0
Codecs	1935.1	5	387.015	527.63	0
Bitrates	1370.7	3	456.911	622.92	0
Perspective views	3.3	2	1.632	2.23	0.1081
Contents*Codecs	65.7	25	2.627	3.58	0
Contents*Bitrates	45.6	15	3.038	4.14	0
Contents*Perspective views	9.5	10	0.953	1.3	0.224
Codecs*Bitrates	738.4	15	49.225	67.11	0
Codecs*Perspective views	7.3	10	0.729	0.99	0.4459
Bitrates*Perspective views	2.2	6	0.372	0.51	0.8031
Error	5315.6	7247	0.733		

Table 3.8 – Multi-way ANOVA (interaction model) on the raw scores given to reference contents (perspective views only).

	SumSq	DF	MeanSq	F	p-value
Contents	50.6	5	10.1204	22.23	0
Codecs	12.12	5	2.424	5.32	0.0001
Bitrates	0.98	3	0.3269	0.72	0.541
Perspective views	3.79	2	1.8956	4.16	0.0156
Contents*Codecs	14.21	25	0.5686	1.25	0.1822
Contents*Bitrates	2.02	15	0.1348	0.3	0.9959
Contents*Perspective views	5.95	10	0.5947	1.31	0.2203
Codecs*Bitrates	2.23	15	0.1486	0.33	0.9929
Codecs*Perspective views	4.65	10	0.4655	1.02	0.421
Bitrates*Perspective views	1.73	6	0.2888	0.63	0.7028
Error	3298.92	7247	0.4552		

Table 3.9 – Multi-way ANOVA (interaction model) on the raw scores given to test contents (refocused views only).

	SumSq	**DF**	**MeanSq**	**F**	**p-value**
Contents	325.74	5	65.149	82.7	0
Codecs	981.81	5	196.361	249.25	0
Bitrates	602.85	3	200.95	255.07	0
Refocused views	1.15	1	1.149	1.46	0.2273
Contents*Codecs	42.09	25	1.684	2.14	0.0008
Contents*Bitrates	30.82	15	2.055	2.61	0.0006
Contents*Refocused views	22.36	5	4.471	5.68	0
Codecs*Bitrates	393.11	15	26.207	33.27	0
Codecs*Refocused views	7.06	5	1.411	1.79	0.111
Bitrates*Refocused views	2.29	3	0.764	0.97	0.406
Error	3791.72	4813	0.788		

Table 3.10 – Multi-way ANOVA (interaction model) on the raw scores given to reference contents (refocused views only).

	SumSq	**DF**	**MeanSq**	**F**	**p-value**
Contents	43.8	5	8.7598	9.28	0
Codecs	101.15	5	20.2309	21.43	0
Bitrates	2.11	3	0.7045	0.75	0.5244
Refocused views	0.08	1	0.0817	0.09	0.7686
Contents*Codecs	14.35	25	0.5739	0.61	0.9363
Contents*Bitrates	5.21	15	0.3472	0.37	0.9867
Contents*Refocused views	38.01	5	7.6018	8.05	0
Codecs*Bitrates	6.38	15	0.4254	0.45	0.9639
Codecs*Refocused views	7.1	5	1.4209	1.51	0.1846
Bitrates*Refocused views	1.16	3	0.3878	0.41	0.7453
Error	4543.63	4813	0.944		

Table 3.11 – Multi-way ANOVA (interaction model) on the raw scores given to test and reference contents, for all rendered views.

	SumSq	DF	MeanSq	F	p-value
Contents	544.6	5	108.922	89.47	0
Codecs	1151.1	5	230.23	189.12	0
Bitrates	917.8	3	305.921	251.3	0
Views	497.7	4	124.435	102.22	0
Contents*Codecs	71.9	25	2.876	2.36	0.0001
Contents*Bitrates	42.4	15	2.827	2.32	0.0026
Contents*Views	78.4	20	3.922	3.22	0
Codecs*Bitrates	591	15	39.398	32.36	0
Codecs*Views	50.2	20	2.511	2.06	0.0035
Bitrates*Views	15.5	12	1.288	1.06	0.3915
Error	29648.9	24355	1.217		

bitrate (at $p = 0.3915$, the interaction between bitrates and views is not significant).

Results show that, within the same rendering group, different parameters do not lead to significantly different scores, both in test and reference contents. Thus, it is unnecessary to test different rendering parameters within the same group, as it increases the complexity of the test without bringing any added value. On the other hand, different types of rendering, such as perspective and refocused rendering, lead to statistically different results in both test and reference contents. In particular, refocused contents were consistently rated lower than their perspective counterpart. This could suggest that selecting only one of the two types of rendering could lead to biases in the way scores are distributed.

One straightforward conclusion from the analysis reported in this section would be to select the rendering parameters as to have only one view per type of rendering. However, using only one rendering parameter per group could lead to unwanted effects. For example, using only one perspective view to assess the quality of the entire light field content could be a feasible solution if the compression artifacts are homogeneously distributed among the views - that is, if the compression algorithm affects different views in equal measure. If that is not the case, selecting which view should be used in the test may become a delicate task. Indeed, a wrong selection of rendering parameters can favor or penalize certain algorithms or solutions. Moreover, compression solutions might be engineered to offer the best quality for the rendering parameters selected for the test, disregarding the quality of others. Finally, results obtained by assessing only few rendering parameters might be hard to generalize.

In conclusion, although it is theoretically possible to use single-image methodologies to assess light field contents, it is discouraged due to the number of contraindications associated with it, which might lead to biased results. Increasing the number of rendering parameters is not guaranteed to produce corresponding diversity in the scores; thus, its advantages are definitely

outweighted by the increased length and cost it requires. On the other hand, using only a few rendering parameters could be proven ill-advised for certain compression algorithms, and could be susceptible to ad-hoc engineering to achieve the best results at the expenses of the general quality of the content. It is preferable to combine several rendering parameters in one single stimulus to be assessed, for example by employing a pre-recorded animation or by using an interactive setup. We will explore both options in the next section.

3.2 Comparison of passive and interactive methodologies

One of the most exciting properties of light field imaging is the possibility either to visualize the acquired image data directly, or to produce new visual effects (e.g. change of perspective, refocus, change in DOF, etc.) of the captured scene prior to display. However, as we have asserted before, the abundance of rendering modalities for this type of content poses several challenges in assessing its visual quality. In our previous section, we have presented how different variations of rendered views do not always correspond to significantly different ratings. Moreover, we have showed that separate visual effects, such as change of perspective and change of focal point, can lead to important discrepancies in the perceptual quality, even in the absence of compression distorsions.

Considering that evaluating single rendered views implies that a large number of stimuli will be tested, thus increasing the length of the test and causing fatigue on subjects, it is more efficient to employ methodologies that enable global assessment of quality of experience. This can be implemented either with an interactive setup, which would allow to engage with the contents in a flexible and intuitive way, or with an automatic presentation of rendered views in form of an animation, which would show different rendered views of the light field content under test. Both approaches have the benefit of reducing the number of stimuli to be tested, while allowing evaluation of the light field content as a whole.

The automatic presentation has been favored by several publications in literature, which employ it to evaluate the quality of light field contents in the presence of various artifacts. Paudyal et al. study the effect of watermarking on light field rendering, performing a test using a circular animation of perspective views [Paudyal et al., 2015]. Perra et al. analyse how light field subsampling affects the perceived quality of refocused views, which are presented in an animated fashion [Perra et al., 2018]. Battisti et al. compare several visualization techniques, as well as different framerates, for light field quality assessment, concluding that horizontal scan is to be preferred [Battisti et al., 2018]. The interactive assessment, on the other hand, is favoured by Shi et al., who implement an interactive framework to perform visual quality assessment of light field images on 3D screens [Shi et al., 2018].

Advantages and drawbacks can be drawn for both approaches. As we have mentioned in the introduction to this chapter, interactivity represents the most natural way for users to engage with the content, whereas a passive approach such as the animated presentation reduces the intriguing features of light field photography to a traditional video content. On the other hand,

(a) Bikes (b) Stone_Pillars_Outside (c) Fountain_&_Vincent_2 (d) Friends_1

Figure 3.5 – Central perspective image from each content used in our experiment. ©2017 IEEE

Table 3.12 – Values of refocusing slopes for each content. ©2017 IEEE

	Slopes										
Content	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Bikes	-10	-8	-6	-4	-2	0	2	4	6	8	10
Stone_Pillars_Outside	-10	-8	-6	-4	-2	0	2	4	6	8	10
Fountain_&_Vincent_2	-10	-8	-6	-4	-2	0	2	4	6	8	10
Friends_1	-5	-4	-3	-2	-1	0	1	2	3	4	5

Table 3.13 – Test environments and specifications. ©2017 IEEE

Approach	**Environment**	**No. subjects**	**Methodology**	**No. perspective views**	**No. refocused views**	**fps**	**Median age**
Interactive	Controlled lab setting	24	DSIS	169	11	-	25
Passive	Semi-controlled crowdsourcing	24	DSIS	97	11	30	22

with the former approach there is no guarantee that every user will experience the content in the same, reproducible way; as every subject freely chooses what rendered image to visualize and for how long, more variability is added in the test. Conversely, the passive approach ensures that every subject undergoes the same procedure, leading to a uniform assessment.

In this section, we compare results of subjective assessments of visual quality obtained by using two methodologies, one that enforces interaction with the content, and one that favors an automated presentation. For the first methodology, a controlled lab environment was adopted, while for the second methodology, due to time and costs constraints, a crowdsourcing tool was deployed. In order to perform a meaningful comparison between the two methodologies, five compression solutions were selected from the literature. For a comprehensive overview of each solution and a thorough comparison through both objective and subjective means, we refer the readers to Chapter 7.

Table 3.14 – Summary of compression schemes. ©2017 IEEE

Proponent	Description
P01	Lenslet image compressed using HEVC intra (software x265).
P02	Lenslet image compressed using HEVC intra with LLE and SS (software HM-14.0) [Monteiro et al., 2016].
P03	Lenslet image compressed using intermediate transformation to perspective views and HEVC (software JEM 2.0) [Liu et al., 2016].
P04	Chroma subsampling of the lenslet image and compression of perspective views through pseudo-temporal sequence using HEVC (software x265).
P05	Compression of perspective views through pseudo-temporal sequence using HEVC (software x265).

3.2.1 Experimental test design

This section describes how the subjective evaluations were designed. More specifically, the creation of the stimuli for both tests is outlined. A description of the interactive subjective methodology, along with the testing environment, is presented. Then, the passive subjective methodology is described in details. A summary of the specifications for the two methodologies can be found in Table 3.13.

Data preparation

Four light field imageswere selected from the aforementioned EPFL light field image dataset, namely contents *Bikes*, *Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* [Řeřábek and Ebrahimi, 2016]. Thumbnails for each content are depicted in Figure 3.5. Following ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012], the images were carefully selected in order to provide a wide range of scenarios, including details that would prove critical for the compression algorithms.

The lenslet images were processed using the Light Field MATLAB toolbox [Dansereau et al., 2013, 2015] to obtain the collection of perspective views needed for the subjective tests. Additionally, eleven refocused images were created for each content, using a modified version of the toolbox function *LFFiltShiftSum*. For our tests, it was decided to sum images from index 3 to index 13 (11×11 images) to have a larger DOF than that obtained by shifting and summing all of the perspective views. The values of the slopes used to shift the perspective views are summarized in Table 3.12. The slopes were selected to assure gradual transition between refocusing on the foreground and on the background with respect to semantically relevant objects in each content.

The uncompressed reference was obtained by preprocessing the raw sensor data through devignetting, demosaicing, clipping to 8 bits, transforming to a collection of perspective views

and applying color and gamma corrections. The reference was obtained from the lenslet image in RGB 444, without any chroma subsampling. This reference was selected to have a proper comparison with acquisition data obtained with minimal pre-processing. For this reason, chroma subsampling was not applied on the reference, since it alters the data.

Five compression algorithms were used to create the data to evaluate the two methodologies. Three anchors were created by the authors using HEVC encoding (x265 implementation), whereas two others were taken from literature [Monteiro et al., 2016, Liu et al., 2016]. Each compression scheme was given a label for easier identification. A summary of the compression schemes can be found in Table 3.14. The compression algorithms were evaluated on four bitrates (corresponding to four compression ratios), namely $R1 = 1$ bpp $(10:1)$, $R2 = 0.5$ bpp $(20:1)$, $R3 = 0.25$ bpp $(40:1)$, $R4 = 0.1$ bpp $(100:1)$. The compression ratios were computed as ratios between the size of the uncompressed raw images in 10bit precision and the size of the compressed bitstreams.

Interactive methodology

To perform the interactive visual assessment, a methodology for evaluation of plenoptic content was selected [Viola et al., 2016b]. The methodology is based on DSIS [ITU-R BT.500-13, 2012].

Participants were asked to interact with the light field images and rate the level of impairments of the test light field image with respect to the reference, on a scale from 1 (*Very annoying*) to 5 (*Imperceptible*). Each light field image was presented together with the uncompressed reference in a side-by-side fashion. The position of the reference was set to either left or right for each experiment, and participants were informed about its location on the screen. For each stimulus, the central perspective view image from the light field image was displayed. By clicking inside the displayed image and dragging the mouse, the other perspective views from the light field image were accessed and displayed. Each image was displayed in its native resolution of 625×434 pixels. A total of 13×13 perspective views were accessible. The refocused views were accessible through a slider shown at the bottom of each stimulus.

To avoid the involuntary influence of external factors and to ensure the reproducibility of results, the laboratory for subjective video quality assessment was set up according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. Professional Eizo ColorEdge CG301W 30-inch monitors with native resolution of 2560×1600 pixels were used for the tests. The monitors were calibrated using an i1Display Pro color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level of 0.2 cd/m^2. The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. The illumination level measured on the screens was 15 lux. The distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content, conforming to requirements in ITU-R Recommendation

Table 3.15 – Selected settings for AVC coder for passive methodology. ©2017 IEEE

-r 30 -s <*size*> -f rawvideo -pix_fmt yuv420p -i <*input*> -c:v libx264 -profile:v high -x264opts no-scenecut:no-deblock:pass=1 -b:v 8M tmp.mp4
-r 30 -s <*size*> -f rawvideo -pix_fmt yuv420p -i <*input*> -c:v libx264 -profile:v high -x264opts no-scenecut:no-deblock:pass=2 -b:v 8M <*output*>

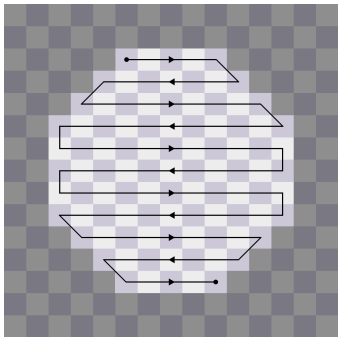


Figure 3.6 – Ordering of the views for animation for passive methodology. ©2017 IEEE

BT.2022 [ITU-R BT.2022, 2012].

Before the experiments, a training session was organized to allow participants to get familiar with artefacts and distorsions in the test images. Five training samples were manually selected by expert viewers. The training samples were created by compressing other content on various bitrates. The content used for the training was selected from the same light field image database used for the test images [Řeřábek and Ebrahimi, 2016]. The training samples were presented along with the uncompressed reference, exactly as they were shown in the tests.

The experiment was split into two sessions. In each session, 40 stimuli were shown side by side with the uncompressed reference, corresponding to approximately 20 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all the sessions, thus evaluating the entire set of stimuli. A break of ten minutes was enforced between the sessions to avoid fatigue. Before the test, one dummy sample was inserted to ease the participants into the task. The resulting scores from dummy stimuli were not included in the results.

A total of 24 subjects (19 males and 5 females) participated in the experiment, for a total of 24 scores per stimulus. Subjects were between 18 and 35 years old, with an average of 24.79 and a median of 25 years of age. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts.

Passive methodology

The passive visual assessment of quality was carried on using a methodology based on DSIS [ITU-R BT.500-13, 2012]. To perform the tests, the QualityCrowd 2 framework [Keimel et al., 2012] was used. However, it should be noted that all the participants performed the tests in the same environment at the same time, with equal lighting conditions, using the same display model and the same screen resolution. The participants were shown the light field content as a video sequence navigating between the perspective and the refocused views. Each stimulus was displayed alongside with the uncompressed reference, in a side by side fashion. The subjects knew in advance on which side of the screen the reference was displayed.

Due to distortions caused by the lenslet structure, several perspective views presented artefacts independent from the coding procedure, and thus had to be discarded. Only a subset of 97 out of 225 perspective views was chosen to be displayed, in order not to affect the rating. Ten perspective views per second were displayed, to ensure a smooth transition of the different views. The perspective views were accessed from top to bottom and from left to right and right to left in alternate order (see Figure 3.6). At the end of the animation of the perspective views, the eleven refocused views were displayed with a framerate of four refocused views per second, going from foreground to background and from background to foreground. The animation setup was chosen and validated by expert viewers in order to mimic the parallax effect, as well as to mimic the refocusing effect that occurs when trying to change the focal point. The total length of the animation for each stimulus was 14 seconds. Since there is no browser video plugin capable of reliable real-time decoding and displaying for HEVC, the animations were encoded with AVC. A two-pass encoding was used and the deblocking filter was disabled to ensure transparency and to preserve the original blockiness artefacts when encoded at low bit rates. Expert viewing session conducted prior to the main subjective assessment concluded that the AVC video encoding was visually lossless, and thus would not influence in any way the final scoring. Selected settings for AVC coder are summarised in Table 3.15.

Test subjects were asked to rate the level of impairment of the test stimuli when compared to the uncompressed references. The rating was performed on a scale from 1 (Very annoying) to 5 (Imperceptible). Before the experiment, a training session was organized to allow participants to get familiar with artefacts and distorsions in the test images. Five training samples among the compressed stimuli were manually selected by expert viewers. To help subjects localize and identify compression artefacts in the fast-paced video, the same content used in the test was selected for the training. The training samples were presented along with the uncompressed reference, exactly as they were shown in the test.

The experiment was split into two sessions. In each session, 40 stimuli were shown side by side with the uncompressed reference, corresponding to approximately 20 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all the sessions, thus evaluating the entire set of stimuli. A break of ten minutes was enforced between the sessions to avoid fatigue.

A total of 24 subjects (22 males and 2 females) participated in the experiment, for a total of 24 scores per stimulus. Subjects were between 18 and 35 years old, with an average of 22.79 and a median of 22 years of age.

3.2.2 Statistical analysis

Outlier detection and removal was performed on the results, independently for each methodology, according to the ITU Recommendations [ITU-R BT.500-13, 2012]. One outlier was detected in results obtained using the interactive methodology, whereas no outlier was found in the results from the passive methodology. This led to 23 scores per stimulus for the first method, and 24 scores per stimulus for the second. After outlier removal, the MOS was computed for each stimulus, independently for each methodology. The corresponding 95% CIs were computed assuming a Student's t-distribution.

Following the ITU Recommendations [ITU-T P.1401, 2012], several fittings were applied to the MOS values from the two different methodologies. In particular, first order and third order fittings were used to compare the MOS values. Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), Spearman's Rank Correlation Coefficient (SRCC) and Outlier Ratio (OR) were computed for accuracy, linearity, monotonicity and consistency, respectively.

A multiple comparison test was performed at a 5% significance level on the raw scores, to determine, for each stimulus, whether the MOS values obtained with the two methodologies were significantly different, and the percentage of correct estimation, underestimation and overestimation were computed. Additionally, the classification errors were computed using the same multiple comparison test to see if the results obtained with the two methodologies lead, for each pair of stimuli, to the same conclusions [ITU-T J.149, 2004]. In this case, three types of error can be distinguished: false ranking, false differentiation and false tie. False ranking is the most offensive error, and occurs when the first methodology says that situation i is better than situation j, whereas the second methodology says the opposite. False differentiation occurs when the first methodology says that situation i and j are different, whereas the second methodology says they are the same. False tie occurs when the first methodology says two situations are the same, whereas the second methodology says they are different.

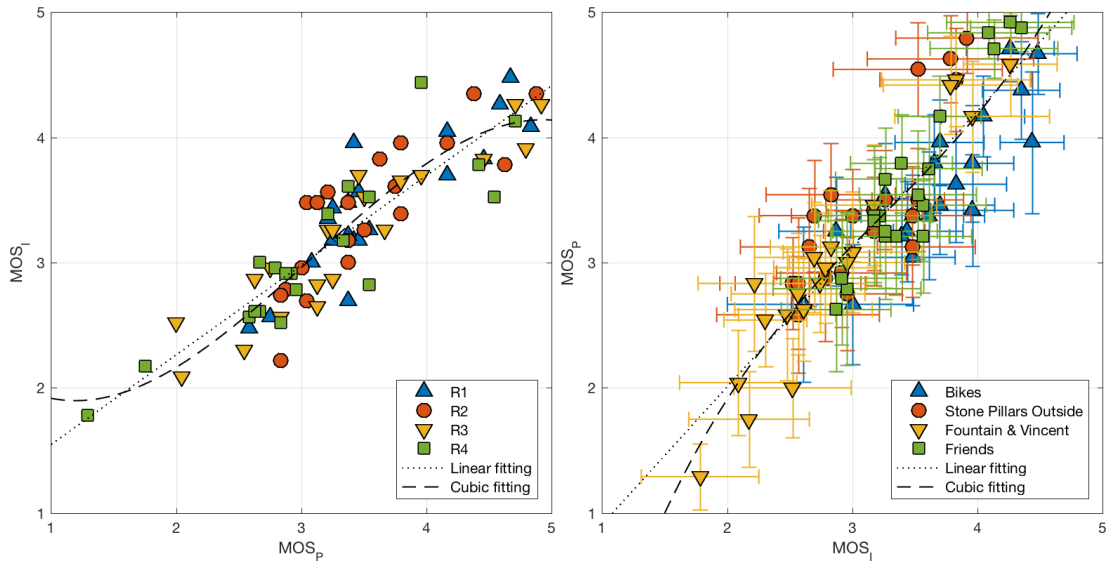
Finally, one-way and multi-way ANOVA tests were performed to assess the influence of the methodology on the results, and in particular whether the two methodologies lead to significantly different results.

3.2.3 Results and discussion

Figure 3.7 shows the scatter plots comparing the MOS values obtained with the two tested methodologies. On the right, the horizontal and vertical bars represent the CIs corresponding to results obtained with interactive and passive methodologies, here denominated I and P, respectively. To improve visualization, the points are colored based on compression ratio or

Table 3.16 – Performance indexes. ©2017 IEEE

	$[\widehat{MOS_P}, MOS_I]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Correct Decision**	**False Ranking**	**False Differentiation**	**False Tie**
No fitting	0.8878	0.8876	0.3791	3.75%	100%	84.56%	0.00%	13.04%	2.41%
Linear fitting	0.8878	0.8876	0.2797	0.00%	100%	89.37%	0.00%	3.26%	7.37%
Cubic fitting	0.8957	0.8876	0.2708	0.00%	100%	88.80%	0.00%	0.82%	10.38%
	$[\widehat{MOS_I}, MOS_P]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Correct Decision**	**False Ranking**	**False Differentiation**	**False Tie**
No fitting	0.8878	0.8876	0.3791	3.75%	100%	84.56%	0.00%	2.41%	13.04%
Linear fitting	0.8878	0.8876	0.3468	0.00%	100%	86.84%	0.00%	3.26%	9.91%
Cubic fitting	0.8895	0.8876	0.3444	0.00%	100%	89.97%	0.00%	6.42%	3.61%



(a) MOS_I as function of MOS_P.

(b) MOS_P as function of MOS_I.

Figure 3.7 – Comparison of MOS values obtained with the different methodologies, along with linear and cubic fittings. Points are differentiated by compression ratio (a) and by content (b). ©2017 IEEE

content. Linear and cubic regressions are shown for both comparisons. Table 3.16 shows the performance indexes computed on the data. The indexes are computed on the data pairs $[\widehat{MOS_A}, MOS_B]$ where $A, B = I, P$. $\widehat{MOS_A}$ are the MOS scores obtained with methodology A with no fitting, linear fitting and cubic fitting, and MOS_B are the MOS scores obtained with methodology B.

Ideally, a 45° line would indicate that the two methodologies give the same MOS values for the same condition. However, as it is visible in Figure 3.7, the points are not aligned along the $y = x$ line. In particular, linear regression performed on MOS_P has a slope of 0.716 and an intercept of 0.832, which indicates that, on average, for the same stimulus subjects gave a higher rating when presented with passive methodology as opposed to interactive methodology. This is confirmed by the results of boxplot analysis on the two methodologies, which shows that, on average, results obtained with the passive methodologies tend to have higher ratings. This tendency can be explained considering that viewers are presented with a carefully selected subset of perspective views in the passive experiments, which are less prone to lenslet-based artefacts, as opposed to the wider number of perspective views subjects can access in the interactive experiments.

Cubic regression has a sigmoid shape in both $\widehat{MOS_P}$ and $\widehat{MOS_I}$, as confirmed by values obtained performing PCC and SRCC, which indicate a strong but not perfect linear correlation. Low values of RMSE and OR confirm the correlation between the two methodologies.

Furthermore, there is no over- or under-estimation, as proven by correct estimation being 100%, which indicates that, for the same stimulus, there is no statistically significant difference between the scores obtained with one or the other methodology.

One-way ANOVA performed on stimuli grouped only by methodology shows that results obtained with the two methodologies, although highly correlated, are statistically significantly different ($p = 0.0005$). To further investigate the influence of the coding parameters on the scores, we performed multi-way ANOVA on the results, separately for different compression ratios, contents and codecs, respectively. Results show that, for compression ratios $R2$ and $R4$, the two methodologies are statistically equivalent at 5% significance level, whereas for the remaining compression ratios they are statistically significantly different ($p = 0.008$ and $p = 0.0046$ for $R1$ and $R3$, respectively). Thus, the impact of the methodology on the resulting score does not appear to be driven by the compression ratio. For content *Bikes*, the two methodology are statistically equivalent, whereas for the remaining contents the two methodologies are statistically different ($p = 0$, $p = 0.044$ and $p = 0.0131$ for contents *Stone_Pillars_Outside, Fountain_&_Vincent_2* and *Friends_1*, respectively), meaning that the choice of methodology could have an impact on how the contents are rated. Finally, multi-way ANOVA analysis on different codecs shows that $P5$ is the only codec for which the two methodologies provide statistically different results ($p = 0$).

The classification errors show that there is no false ranking, the most offensive error. However, results from false differentiation performed on $[\widehat{MOS_P}, MOS_I]$ with no fitting show that, on 13.04% of cases, passive methodology considers two stimuli as being statistically significantly different, whereas the interactive methodology does not differentiate them. The percentage thus shows that the passive methodology has more discriminating power when compared to the interactive methodology. This is confirmed by comparing the CIs obtained with the two methodologies: on average, CIs obtained with passive methodology are 8.66% smaller. In other words, the standard error obtained with interactive methodology on 23 subjects would be equivalent to the standard error obtained with passive methodology on 20.13 subjects. Conversely, when using the interactive methodology, 27.42 subjects would be needed to obtain the same standard error provided by the passive methodology on 24 subjects.

It should be noted that, whereas the interactive evaluation has been conducted in a lab setting compliant with the guidelines set by ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012], the passive evaluation has been carried out using crowdsourcing, which is usually associated with less reliable scores. However, several studies have proven the efficacy of crowdsourcing-based tests [Ribeiro et al., 2011, Saupe et al., 2016]. Moreover, while crowdsourcing is usually linked to larger standard errors, due to variability of conditions, the opposite has been observed in our experiment. It shows that the passive approach contributed to lower the variance of the scores, in spite of the impact crowdsourcing might have in increasing the variance of the results.

3.3 Conclusions

In this chapter we presented several methodologies for visual quality assessment of light field contents using image-based rendering. We first showed that deploying single-image evaluation methodologies, which considerably increases the number of stimuli to be assessed, does not lead to a diversification in scores, and if not properly organised can lead to biased ratings. Concluding that combining several renderings in one stimulus would be preferable, we then proceeded to compare two methodologies for light field contents: one that allows interaction with the rendered images to favor engagement with the contents, and another that uses a pre-recorded animation of the rendered views to ensure reproducibility. Results showed that, although the two approaches are strongly correlated, they are not statistically equivalent. In particular, the latter leads to smaller CIs and thus ensures more discriminative power among the tested solutions.

Our contributions in this chapter can be recapped as follows:

- We present three different subjective methodologies for visual quality assessment of light field contents. We present advantages and drawback for each of them before proceeding to analyse them in detail.
- We perform an in-depth analysis of single-image assessment for light field contents using widely-used statistical tools. In particular, we test whether different types of rendering (in our case, change of perspective and change of focal point) lead to statistically different scores, and if testing a variety of rendering parameters is advisable. We prove that, within each type of rendering, no statistical difference can be discerned. Thus, it is sufficient to evaluate only one rendered view from each group, as the scores are statistically equivalent. However, between different types of rendering statistically significant differences can be found. We underline that such differences are present in both test and reference contents, thus they cannot be attributed to the effect of compression artifacts.
- We conclude that single-image assessment is ill-suited for the evaluation of light field contents, as certain types of rendering can affect the final ratings. As multiple renderings of the same types were proven redundant, only one view from each category of rendering should be selected to optimize the length of the test; however, choosing which view should be evaluated can be delicate, as separate views could be affected differently by coding artifacts. Thus, aggregating multiple renderings in the same stimulus seem to be a more promising scenario.
- We present two methodologies for light field contents that combine several rendered views in order to reduce the number of stimuli to be tested. The first allows the users to interact with the light field content, changing the rendering parameters as they please. The second presents them with an animated sequence of several rendered views, which can be passively visualised and scored as a traditional video content. We perform a test

to compare the two methodologies, showing that they are highly correlated, although not statistically equivalent, and lead to similar ratings. However, we found that the interactive approach leads to larger CIs in the corresponding scores, due to lack of control over the number of views that each participant visualises. Conversely, the passive approach, although conducted in a less controlled environment, showed a significant reduction in CIs, and thus an increased discriminative power.

Our main recommendation would be to prefer a passive methodology when being able to discriminate among several solutions is required. However, we want to remark that interaction is a very desirable feature in light field quality assessment. Future design of evaluation methodologies for light field contents should consider improving consistency for interactive testings, for example by merging it with a passive approach, to ensure the same visualization experience for all users, while still enabling interaction with light field contents.

4 Analysis of interaction patterns in light field quality evaluation

Image-based rendering of light field contents offers the possibility of changing the appereance of the scene as it was captured from the acquisition device: the perspective can be changed, the focal plane can be moved, shrinked or extended, and depth planes can be bypassed, among other visual effects. Such a rich scene representation poses new challenges in assessing the visual quality of the content.

Evaluation of visual quality and user experience plays a fundamental role in designing effective and efficient compression solutions, as well as new rendering techniques. However, only few publications are focused on discussion about subjective methodologies for light field content.

Some preliminary work has been performed on subjectively assess the quality of light field contents on light field displays. Spatial resolution of back-projected displays has been investigated by Kovacs et al. [Kovács et al., 2014]. In particular, the authors investigate how viewing angle affects the perception of spatial resolution, along with the role played by motion parallax. Darukumalli et al. investigate the relationship between zooming levels, region of interest and subjective quality of light field contents, using ACR and DQR [Darukumalli et al., 2016]. Kara et al. analyse the impact of angular resolution on the perception of light field content, first in a

free movement scenario, and then with fixed observer position, using ACR [Kara et al., 2016, 2017b].

Light field displays, although commercially available, have not yet seen a widespread success, mainly due to their cost and the requirements for room setup. On the other hand, image-based rendering represents a way to engage with light field contents using legacy 2D displays, which are widely available to consumers. However, few publications have been focusing on QoE for image-based rendered light field contents.

The most natural way of experiencing the capabilities of light field image-based rendering is by enabling interaction with the content in a real-time framework. The possibility of interaction with the acquired content by changing the appearance of the scene has already been proven as a desirable feature in mainstream social media, such as Instagram, Snapchat and Facebook. However, as we showed in Chapter 3, the interactive methodology had less discriminative power when compared to a passive approach, and leads in general to larger CIs.

In order to retain the discriminative power witnessed in passive approaches, while maintaining the interactive features that define light field contents, a thorough analysis of user behaviour when engaging with such content is needed. Although the subject has been investigated in relation to stereoscopic or light field displays, no work has been presented on analysing patterns in user interaction for image-based rendering.

Three main areas of impact can be identified regarding the analysis of user behaviour:

- **Perceptual coding**. Tracking user behaviour with light field contents leads to statistically accurate knowledge on which rendered views composing the light field content are perceptually meaningful and are more frequently accessed. This information can be used when designing new compression algorithms based on perceptual features.
- **Weighting of objective quality metrics**. Currently available objective quality metrics for light field contents, such as those used in ICIP 2017 Grand Challenge on Light Field Coding [Viola and Ebrahimi, 2018a], give the same weight to all rendered views. A weighted average based on relative frequency of access for every view could be more effective in predicting subjective scores for new light field contents.
- **Design of subjective methodologies**. Interactive subjective methodologies have been shown to be less discriminative than passive subjective methodologies, one of the main reasons being that not all subjects are visualizing strictly the same content. Analysing how users interact with the content will help designing new tests that incorporate user behaviour, thus bridging the gap between interactive and passive subjective methodologies.

In this chapter, we present some preliminary results on analysis of user interaction for light field visual quality assessment. To do so, we created a new open-source software for light field

(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside (d) Fountain_&_Vincent_2 (e) Friends_1

Figure 4.1 – Central perspective view from each content used in the test.

quality assessment that allows users to freely engage with the content (see Annex B). The user interaction data is recorded through a validating experiment, and it is subsequently analysed in order to extract meaningful patterns in user behaviour.

4.1 Experiment design

4.1.1 Dataset preparation and description

For the experiment five light field contents, acquired with a Lytro Illum camera, were chosen from EPFL light field image dataset [Řeřábek and Ebrahimi, 2016]. More specifically, contents *Bikes*, *Danger_de_Mort Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* were used in our experiments. Figure 4.1 depicts the thumbnail of the central perspective views from each content.

Each 10bit raw lenslet image was devignetted, demosaiced, and transformed into an light field data structure of perspective views using the Light Field toolbox v0.4 [Dansereau et al., 2013, 2015]. A total of 15 × 15 perspective views were created from the lenslet image, each having a resolution of 625 × 434 pixels. The perspective views were subsequentially saved in *ppm* file format, with 10 bits per color channel, to serve as reference.

Two codecs were adapted for compression of light field evaluated in the test. Both codecs perform the compression on the perspective views, which were preemptively ordered in a pseudo-temporal sequence. To be used as input for the compression algorithms, the perspective images were padded with black pixels, converted to YCbCr format and downsampled from 444 to 422, 10-bit depth. Then, they were arranged in a pseudo-temporal arrangement (see Figure 4.2) and saved in *yuv* file format.

The first codec that was used to compress the pseudo-temporal sequence consisted in HEVC Main10 profile. The software x265 was used to compress the sequence[1]. The full command line used can be found in Table 4.1. The QPs were chosen to match the desired compression

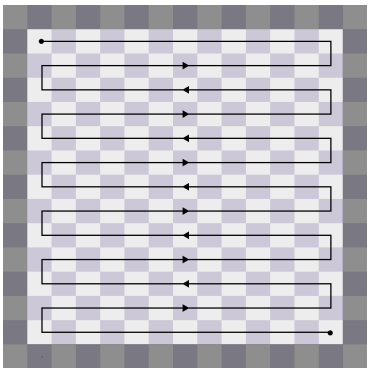


Figure 4.2 – Order of perspective views for pseudo-temporal sequence used for coding.

Table 4.1 – Selected settings for x265 Main10 coder.

--input < *Input* > --input-depth 10 --input-csp i422 --fps 30 --input-res < *Width* > × < *Height* > --output < *Output* > --output-depth 10 --profile main422-10 --crf < *QP* >

Table 4.2 – QP chosen to encode all contents with HEVC.

Content	**R1**	**R2**	**R3**	**R4**
Bikes	13	24	33	44
Danger_de_Mort	15	26	35	43
Stone_Pillars_Outside	14	23	30	40
Fountain_&_Vincent_2	14	24	32	43
Friends_1	12	21	29	40

Table 4.3 – Selected settings for VP9 coder.

--i422 --input-bit-depth=10 --profile=3 -w < *Width* > -h < *Height* > --target-bitrate=< *bitrate*> --cq-level=0 --bit-depth=10 --codec=vp9 --fps=30000/1000 --best -o < *Output* > < *Input* >

ratios. Table 4.2 summarizes the values of different QP used in the test.

As a second codec, VP9 was used to compress the pseudo-temporal sequence[2]. The full command line used can be found in Table 4.3. The target bitrate was chosen to match the corresponding compression ratios defined below.

The test light field content was displayed together with the uncompressed reference in a side-by-side fashion, using the proposed framework. Due to distorsions naturally occurring in lenslet-based light field contents, some of the perspective views were deemed not suitable

[1] https://www.videolan.org/developers/x265.html
[2] https://www.webmproject.org/vp9/

Table 4.4 – Values of refocusing slope for each content.

	Slopes										
Content	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Bikes	-10	-8	-6	-4	-2	0	2	4	6	8	10
Danger_de_Mort	-10	-8	-6	-4	-2	0	2	4	6	8	10
Stone_Pillars_Outside	-10	-8	-6	-4	-2	0	2	4	6	8	10
Fountain_&_Vincent_2	-10	-8	-6	-4	-2	0	2	4	6	8	10
Friends_1	-5	-4	-3	-2	-1	0	1	2	3	4	5

for visualization, since they would negatively bias subjects. Hence, only the central 9×9 perspective views out of the 15×15 views were selected for the test. Both reference and test contents were converted from *ppm* file format in 10 bits to *png* file format in 8 bits, due to limitations of the display and the software.

For each stimulus, the central perspective view from the light field data structure was initially displayed. By clicking inside either test or reference displayed image and dragging the mouse, the other perspective views from the data structure were accessed and displayed. Each image was displayed in its native resolution of 625×434 pixels. Additionally, eleven refocused images of the central perspective view were created for each content, using a modified version of the toolbox function *LFFiltShiftSum* that allows to change the DOF. For the test, it was chosen to sum images from index 3 to index 13 (11×11 images) to have a DOF that is not too narrow, while still showing the effects of refocusing. The values of the slopes are summarized in Table 4.4. The refocused images were accessible through a slider shown between test and reference. Additionally, users could access the refocused images by double clicking on the point of the image they wished to see in focus. The slopes were selected so as to assure gradual transition between refocusing on the foreground and on the background with respect to semantically relevant objects in each content.

The codecs were evaluated on four bitrates, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, $R4 = 0.005$ bpp. The compression ratios are computed as ratios between the size of the uncompressed raw images in 10-bit precision ($5368 \times 7728 \times 10$ bits = 414839040 bits = 10 bpp) and the size of the compressed bitstream.

4.1.2 Testing environment

To avoid the involuntary influence of external factors and to ensure the reproducibility of results, the laboratory for subjective video quality assessment was set up according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. A Samsung SyncMaster 2443 24-inch monitor with native resolution of 1920×1200 pixels was used for the test. The monitor was calibrated using an i1Display Pro color calibration device according to the following profile: sRGB Gamut, D65 white point, 120cd/m^2 brightness, and minimum black level of 0.2 cd/m^2.

Table 4.5 – Test environments and specifications.

Approach	Environment	No. contents	No. codecs	No. bitrates	No. subjects	Methodology	No. persp. views	No. refoc. views
Interactive	Controlled lab setting	5	2	4	23	DSIS	81	11

The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. The illumination level measured on the screens was 15 lux. The distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content, conforming to requirements in ITU-R Recommendation BT.2022 [ITU-R BT.2022, 2012].

4.1.3 Test methodology and planning

The selected methodology was based on DSIS [ITU-R BT.500-13, 2012]. The participants were asked to interact with the light field contents and to rate the level of impairment of the test light field content with respect to the reference, on a scale from 1 (Very annoying) to 5 (Imperceptible). Each content was presented together with the uncompressed reference in a side-by-side fashion, in its native resolution of 625×434 pixels. The position of the reference was fixed for each experiment, and the participants were made aware of its location on the screen (either left or right).

Before the experiments, a training session was organized to allow participants to get familiar with artifacts and distorsions in the test images. Four training samples were manually selected by expert viewers. In order not to influence the results, the training samples were created by compressing other contents on various bitrates. The content used for the training was chosen from the same light field database used for the test images [Řeřábek and Ebrahimi, 2016]. The training samples were presented along with the uncompressed reference, exactly as they were shown in the test.

The test samples were randomly distributed among subjects. The same content was never shown consecutively. Before the test, two dummy samples were inserted to ease the participants into the task. The resulting scores from dummy stimuli were not included in the results.

A total of 23 subjects (11 males and 12 females) participated in the experiment, for a total of 23 scores per stimulus. Subjects were between 18 and 35 years old, with an average of 22.27 and a median of 22.05 years of age. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts. A summary of the specifications of the test can be found in Table 4.5.

4.2 Data processing and statistical analysis

This section describes how data was processed to obtained the results presented in next section. Specifically, subsection 4.2.1 details how subjective scores were processed and analyzed, subsection 4.2.2 enlists the pre-processing and aggregation of user tracking data, while subsection 4.2.3 presents the statistical analysis and cross-correlation of the results.

4.2.1 Subjective scores analysis

Outlier detection was performed according to the guidelines defined in ITU-R recommendation BT.500-13 [ITU-R BT.500-13, 2012]. One outlier was detected and the relative scores were discarded, thus leading to 22 scores per stimulus. The MOS was computed for each coding condition j (i.e., each content, codec and compression ratio) as follows:

$$MOS_j = \frac{1}{N}\sum_{i=1}^{N} m_{i,j}, \tag{4.1}$$

where N is the number of participants and m_{ij} is the score for stimulus j by participant i. The corresponding 95% CIs were computed. To determine whether the results yield statistical significance, a one-sided Welch's test at 5% significance level was performed on the scores, with the following hypotheses:

$$H0 : MOS_A \leq MOS_B$$

$$H1 : MOS_A > MOS_B,$$

in which A and B are the codecs that are being compared. The test was performed for each compression ratio and for each content. If the null hypothesis were to be rejected, then it could be concluded that codec A performed better than codec B for the given content and compression ratio at a 5% significance level.

4.2.2 Tracking information analysis

The total number of seconds spent on each perspective and refocused view are aggregated for each stimulus and for each subjects in matrices $P_{i,j}$ and $R_{i,j}$, respectively:

$$P_{i,j} = \begin{pmatrix} p_{1,1,i,j} & \cdots & p_{1,v,i,j} & \cdots & p_{1,V,i,j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{u,1,i,j} & \cdots & p_{u,v,i,j} & \cdots & p_{u,V,i,j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{U,1,i,j} & \cdots & p_{U,v,i,j} & \cdots & p_{U,V,i,j} \end{pmatrix}, R_{i,j} = \begin{pmatrix} r_{1,i,j} \\ \vdots \\ r_{s,i,j} \\ \vdots \\ r_{S,i,j} \end{pmatrix}, \tag{4.2}$$

where $u = 1,2,\ldots,U$ and $v = 1,2,\ldots,V$ are the indexes of each perspective view, $s = 1,2,\ldots,S$ is the index of each refocused view, $i = 1,2,\ldots,N$ indicates the subject and $j = 1,2,\ldots,M$ indicates the stimulus.

The results were then aggregated to get the total number of seconds each subject spent on each stimulus:

$$\bar{P}_{i,j} = \sum_{u=1}^{U} \sum_{v=1}^{V} p_{u,v,i,j}, \tag{4.3}$$

$$\bar{R}_{i,j} = \sum_{s=1}^{S} r_{s,i,j}, \tag{4.4}$$

$$\bar{T}_{i,j} = \bar{P}_{i,j} + \bar{R}_{i,j}. \tag{4.5}$$

To get the general trend for each stimulus, the mean was computed across all subjects:

$$\hat{P}_j = \frac{1}{N} \sum_{i=1}^{N} \bar{P}_{i,j}, \tag{4.6}$$

$$\hat{R}_j = \frac{1}{N} \sum_{i=1}^{N} \bar{R}_{i,j}, \tag{4.7}$$

$$\hat{T}_j = \frac{1}{N}\sum_{i=1}^{N} \bar{T}_{i,j}. \tag{4.8}$$

4.2.3 Correlation and validation analysis

Statistical analysis was performed on the subjective scores and the results obtained from the tracking of user behaviour, to see whether the results obtained presented some correlation. In particular, statistical analysis was performed between MOS_j, which was used as ground truth, and $\hat{P}_j$, $\hat{R}_j$ and $\hat{T}_j$, for a total of three comparisons. For simplicity, from now on we will refer to $\hat{P}_j$, $\hat{R}_j$ and $\hat{T}_j$ as tracking values.

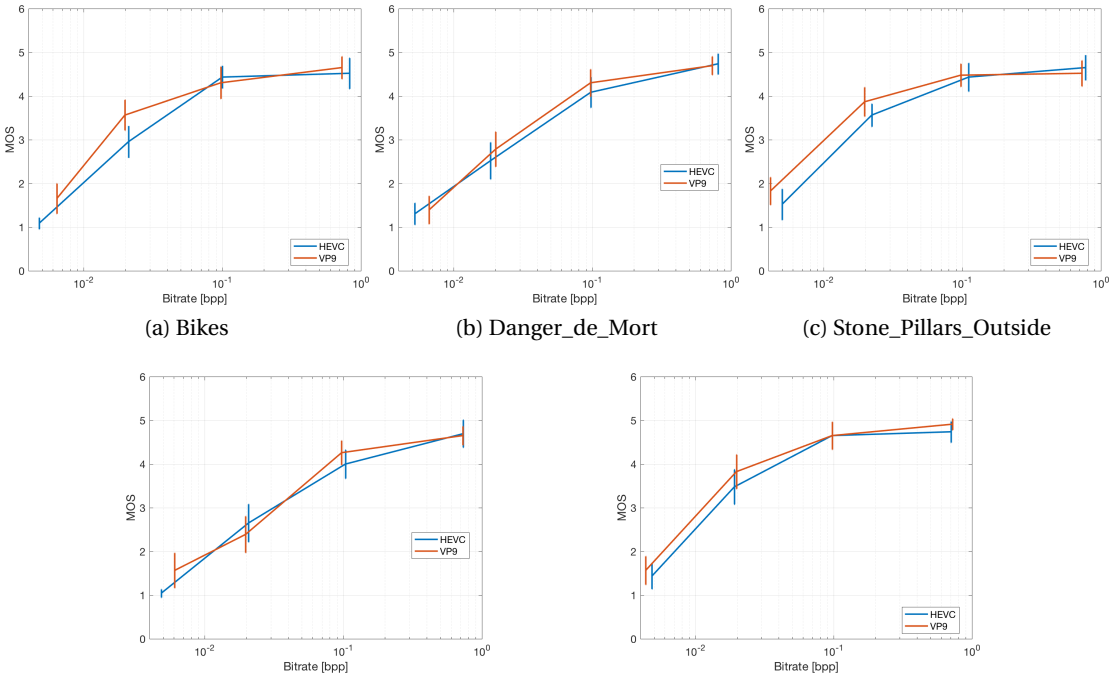
Following the ITU-T Recommendation P.1401 [ITU-T P.1401, 2012], several fittings were applied to the tracking values. In particular, first order and third order fittings were used to compare the values. Absolute prediction error (RMSE), Pearson Correlation Coefficient (PCC), Spearman's Rank Correlation Coefficient (SRCC) and Outlier Ratio (OR) were computed for accuracy, linearity, monotonicity and consistency, respectively.

In order to understand whether the tracking values could effectively be used as predictors for MOS values, estimation and classification errors were computed. A multiple comparison test was performed at a 5% significance level on the raw scores, to determine, for each stimulus, whether the MOS values and the tracking values were significantly different, and the percentage of correct estimation, underestimation and overestimation were computed. Underestimation occurs when the MOS value predicted from the tracking values is significantly lower than the true MOS value. Overestimation, on the other hand, occurs when the predicted MOS value is significantly higher than the true value.

The classification errors were computed using the same multiple comparison test to see if the results lead, for each pair of stimuli, to the same conclusions [ITU-T J.149, 2004]. In this case, three types of errors can be distinguished: false ranking, false differentiation and false tie. False ranking, the most offensive error, occurs when the ground truth says that situation j_1 is better than situation j_2, whereas the predicted MOS obtained from tracking values say the opposite. False differentiation occurs when the true MOS values say that situation j_1 and j_2 are the same, whereas the prediction from tracking results says they are different. False tie occurs when the true MOS scores say two situations are different, whereas the predicted MOS scores say they are the same.

4.3 Results and discussion

This section describes and discusses the results obtained in the evaluation campaign. More specifically, subjective evaluation results are introduced in section 4.3.1. Section 4.3.2 presents the insights provided by tracking of user behaviour, while section 4.3.3 details the correlation



(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside

(d) Fountain_&_Vincent_2 (e) Friends_1

Figure 4.3 – MOS vs bitrate for different contents, with respective CIs. The bitrate is shown in logarithmic scale to improve readability.

and validation results.

4.3.1 Subjective evaluation results

Figure 4.3 shows the MOS against bitrate for all the contents under test, with respective CIs. It can be observed that while the codecs have very similar performance on compression ratio $R1$ and $R2$, some difference can be observed for compression ratios $R3$ and $R4$, where VP9 outperforms HEVC in some of the contents.

The observation is confirmed by the results obtained in Welch's test, summarized in Table 4.6. HEVC is never significantly better than VP9. For compression ratio $R2$, the two codecs are statistically equivalent, whereas VP9 outperforms HEVC on one out of five contents for compression ratio $R1$, two out of five contents for compression ratio $R3$, and three out of five contents for compression ratio $R4$.

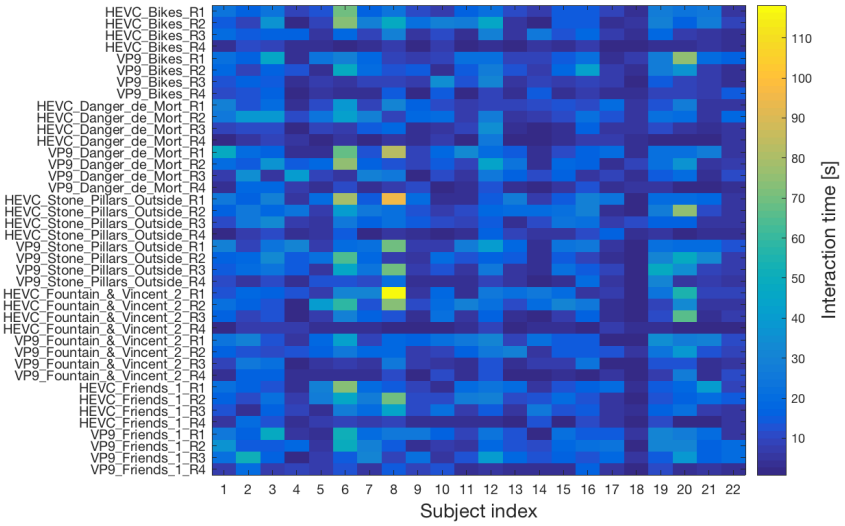


Figure 4.4 – Total interaction time $\bar{T}_{i,j}$ (in seconds) vs stimuli vs subjects.

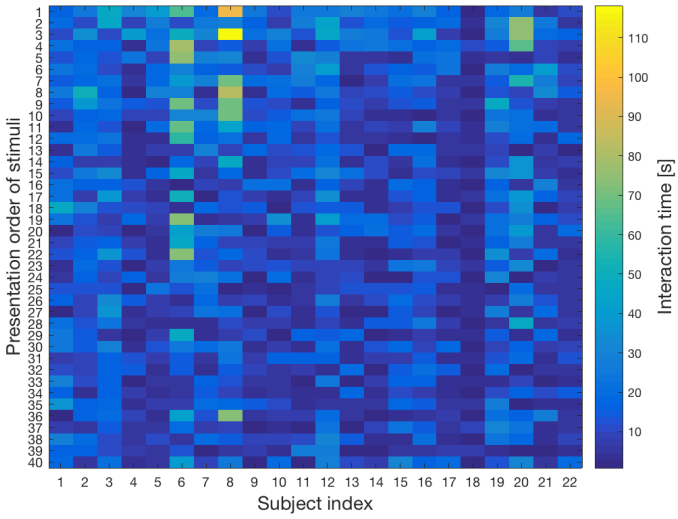


Figure 4.5 – Total interaction time $\bar{T}_{i,j}$ (in seconds) vs order of presentation of the stimuli for each subject.

4.3.2 User tracking results

Figure 4.4 shows the total interaction time $\bar{T}_{i,j}$ for each subject i and each stimulus j. It is noticeable that some users engaged longer with the contents, while others spent on average very little time interacting with the contents (see for example column 6 and 18).

Table 4.6 – Number of contents for which the null hypothesis was rejected, for each compression ratio.

Codec	R1	R2	R3	R4
HEVC	0	0	0	0
VP9	1	0	2	3

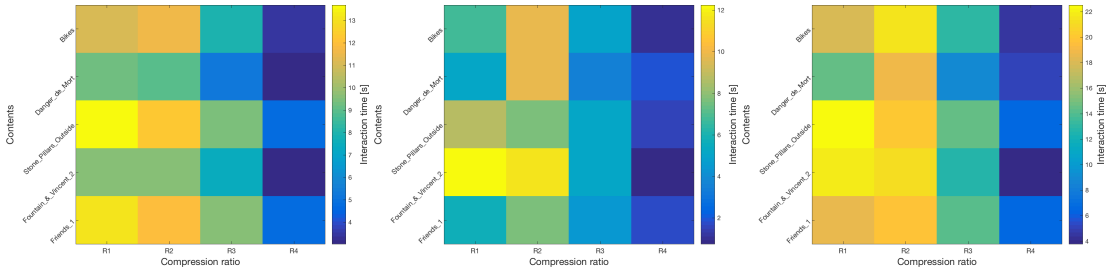
However, a clear trend can be observed among the stimuli. Users seem to spend less time with contents compressed at compression ratio $R4$, as it is visible from the horizontal dark lines in Figure 4.4. In particular, darker lines are present for contents compressed with HEVC at the lowest compression ratio.

To see whether the order of presentation of the stimuli for each subject had any influence on total interaction time $\bar{T}_{i,j}$, the total interaction time was displayed for every subject following the presentation order (see Figure 4.5). Although progressive smaller values of interaction time can be seen as the test progresses, no definite trend can be observed. Hence, it can be concluded that the presentation order had little influence on the total time the users spent interacting with the content. Subjects' boredom and fatigue, along with repetitiveness of the contents, did not have a definite impact on the total time they spent engaging with the stimuli.

Figures 4.6 and 4.7 show the average interaction time $\hat{P}_j$, $\hat{R}_j$ and $\hat{T}_j$, divided by content and by compression ratio, for codec HEVC and VP9, respectively. The results show the trend already observed in Figure 4.4: on average, users tend to interact more with higher bitrates (compression ratios $R1$ and $R2$), whereas for lower bitrates they tend to interact less (compression ratio $R4$). The trend is visible for all type of interactions and for both codecs, although on average people tend to spend more time on codec VP9 for compression ratio $R4$ than they do on codec HEVC.

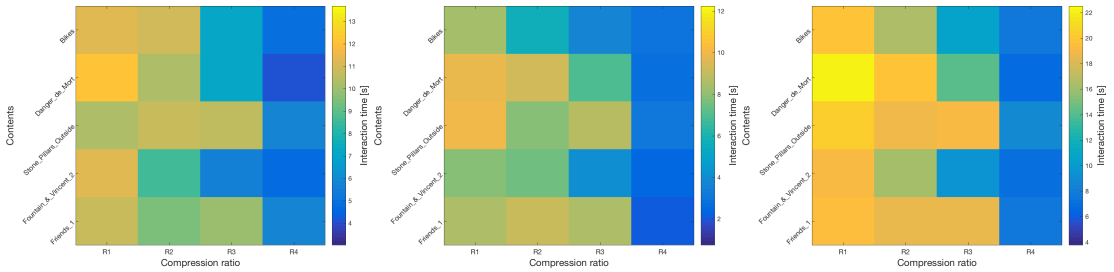
In general, results are more polarized for codec HEVC: users tend to interact more with content compressed with HEVC at higher bitrates with respect to the VP9 counterpart, but they also tend to interact less with content compressed with HEVC at lowest bitrate than they do with the same content compressed with VP9. The average interaction time for codec VP9 is more evenly distributed, although a bitrate-depended trend is still clearly visible (see Figure 4.7).

The average interaction time with perspective views and with refocused views, for both codecs, follows an alternated trend: if users on average spent more time interacting with perspective views for a certain content, they would consequentially spend less time interacting with refocused views. The phenomenon is particularly evident for content *Fountain_&_Vincent_2*. When compressed with codec HEVC at compression ratio $R1$, the content saw an increase in interaction with refocused views, to the detriment of average time spent interacting with perspective views. The opposite behaviour can be observed for codec VP9. However, the total



(a) Interaction time for perspective views. (b) Interaction time for refocused views. (c) Interaction time for all views.

Figure 4.6 – Average interaction time or perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), divided by content and by compression ratio, for codec HEVC.



(a) Interaction time for perspective views. (b) Interaction time for refocused views. (c) Interaction time for all views.

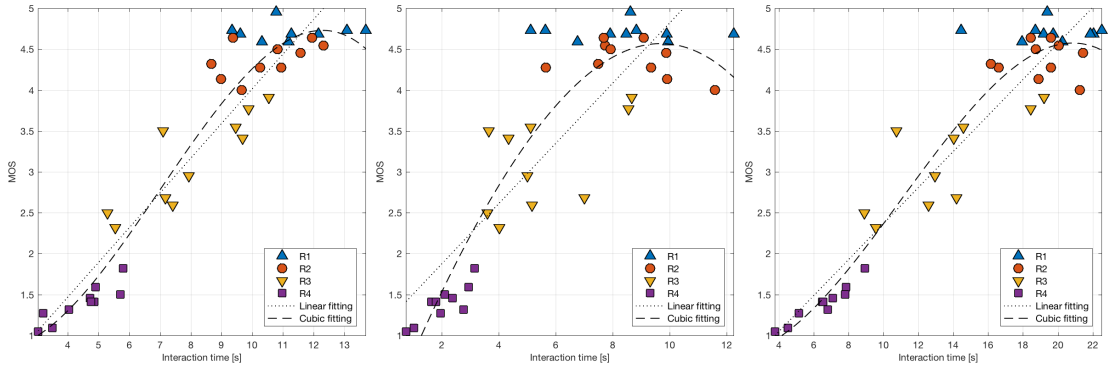
Figure 4.7 – Average interaction time or perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), divided by content and by compression ratio, for codec VP9.

interaction time is quite similar between the two codecs (21.86 against 19.19 seconds).

A trend is also visible regarding the contents, at least when compressed using HEVC (see Figure 4.6). The average interaction time with perspective views was generally higher for contents *Stone_Pillars_Outside* and *Friends_1*, followed by *Bikes*. On the other hand, *Fountain_&_Vincent_2* and *Danger_de_Mort* were, on average, the contents for which the users engaged the least when they needed to interact with perspective views. Interestingly enough, when analysing interaction with refocused views, users engaged more with *Fountain_&_Vincent_2* and *Danger_de_Mort* than they did with the other three contents. As a result, the total average interaction time with all views shows that no visible trend is present for different contents.

4.3.3 Correlation and validation

Figure 4.8 shows the scatter plots comparing the MOS values to the average interaction time for perspective views $\hat{P}_j$, refocused views $\hat{R}_j$, and all views $\hat{T}_j$. To improve visualization, the points were colored based on compression ratio. Figure 4.9 shows the same scatter plots with respective CIs. In this case, points were colored based on the content. Linear and cubic



(a) Interaction time for perspective views vs MOS.

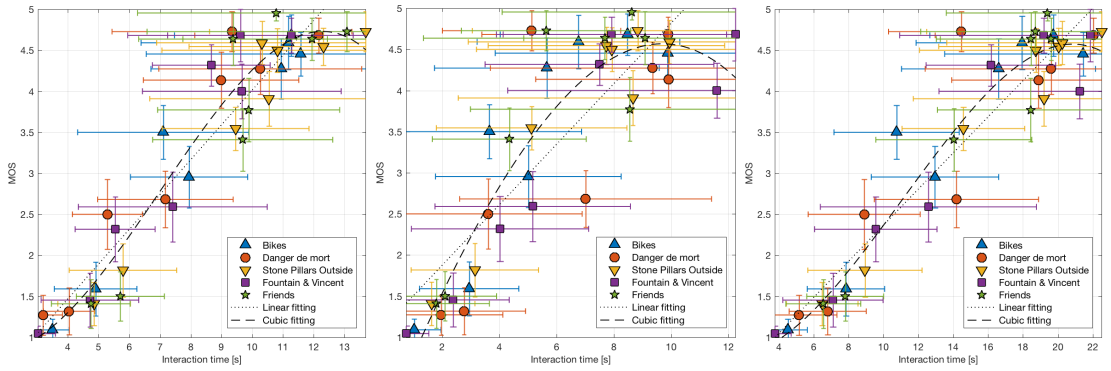
(b) Interaction time for refocused views vs MOS.

(c) Interaction time for all views vs MOS.

Figure 4.8 – Average interaction time for perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), vs MOS. The points are differentiated by compression ratio.

Table 4.7 – Performance indexes.

	$[\widehat{P}, MOS]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Under Estimation**	**Correct Decision**	**False Differentiation**	**False Tie**
Linear fitting	0.9408	0.8704	0.4596	45.00%	100%	0.00%	87.69%	1.41%	10.90%
Cubic fitting	0.9613	0.8511	0.3736	30.00%	97.50%	2.50%	95.38%	4.62%	0.00%
	$[\widehat{R}, MOS]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Under Estimation**	**Correct Decision**	**False Differentiation**	**False Tie**
Linear fitting	0.8543	0.7677	0.7048	82.50%	100%	0.00%	89.74%	1.67%	8.59%
Cubic fitting	0.9161	0.7667	0.5436	57.50%	97.50%	2.50%	97.05%	2.95%	0.00%
	$[\widehat{T}, MOS]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Under Estimation**	**Correct Decision**	**False Differentiation**	**False Tie**
Linear fitting	0.9462	0.8568	0.4387	45.00%	100%	0.00%	90.51%	2.82%	6.67%
Cubic fitting	0.9605	0.8255	0.3774	22.50%	97.50%	2.50%	96.15%	3.85%	0.00%



(a) Interaction time for perspective views vs MOS.

(b) Interaction time for refocused views vs MOS.

(c) Interaction time for all views vs MOS.

Figure 4.9 – Average interaction time for perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), vs MOS, with respective CIs. The points are differentiated by content.

regressions are shown for all comparisons. Table 4.7 shows the performance indexes computed on the data. The indexes are computed on data pairs $[\widehat{X}, MOS]$, in which MOS is the ground truth, and $\widehat{X} = \widehat{P}, \widehat{R}, \widehat{T}$ are the average interaction time results after linear and cubic fitting.

Results from linear and cubic fitting confirm the trend already observed in the previous section. In particular, lower MOS scores are associated with less interaction time on average, whereas longer interaction time is associated with higher MOS scores, for $\hat{P}_j$, $\hat{R}_j$, and $\hat{T}_j$. Results are further confirmed by values obtained performing PCC and SRCC, which show a strong linear correlation.

CIs associated with average interaction time also show that greater variations can be observed for higher MOS scores (see Figure 4.9). As MOS scores decrease, the CIs tend to be smaller as well. Larger CIs can be observed for $\hat{R}_j$ with respect to $\hat{P}_j$ and $\hat{T}_j$.

Although $\widehat{R}$ displays linear correlation with MOS scores, accuracy and consistency are quite low (OR = 82.50% and RMSE = 0.7048 for linear fitting). Indeed, the scatter plot presents a less definite trend with respect to results obtained by $\hat{P}$, especially for compression ratios $R1$ and $R2$ (see Figure 4.8 (a) and (b)). However, adding the results of the interaction with refocused views to the results of the interaction with perspective views helps with both accuracy and consistency (see Figure 4.8 (c) and Table 4.7).

Results from multiple comparison test show that correct estimation is achieved in 100% of cases with linear fitting, and an under estimation of 2.50% is observable when using cubic fitting. Moreover, false ranking, the most offensive error, is never present. Correct decision is achieved on more than 95% of the cases when applying cubic fitting. In general, $\widehat{T}$ has slightly better predictive power than $\widehat{P}$ (i.e., using only perspective views as opposed to using

also refocused views). It also achieves better consistency and accurancy. It can be concluded that, while the average time spent on refocused views alone cannot be used as a predictor for MOS scores, adding it to the average time spent on perspective views improves the correlation results.

Further validation is needed to confirm the predictive power of average interaction time for subjective scores. However, using average interaction time as a predictor for MOS values lays the basis for an implicit quality assessment methodology. One can envision such methodology could be extremely useful in a near future, when plenoptic content will be available on social media and tracking data can be collected anonymously from the natural interaction users have with the content. The tracking data can then be used to predict the quality of the content the users are engaging with, without asking for an explicit score.

4.4 Conclusion

In this chapter we analysed user interaction patterns in light field visual quality assessment. Results showed that clear patterns can be seen in how users interact with the contents, and that strong correlation can be observed between the average interaction time for one content and its corresponding MOS rating. In particular, we showed that the average interaction time can be used to predict the subjective score of light field contents.

The contributions of this chapter can be outlined as follows:

- We design a subjective test for visual quality assessment of light field contents, using a framework to record user interaction, to analyse how user behaviour is affected by compression distortions. More information about the framework can be found in Annex B.
- We proposed a way to aggregate and analyse the recorded user interaction data (tracking information). Specifically, we separated the total number of seconds each subject spent on each perspective and refocused view, per stimulus, to favor analysis of different rendering modalities. We then aggregated the data to get the total number of seconds each subject spent on all the perspective and refocused views, separately for each stimulus. Finally, we provided the average interaction time for perspective and refocused views by computing the mean across the subjects.
- We showed that clear interaction patterns can be extracted from the data. In particular, we showed that the order of presentation does not have a strong influence on the interaction time, and that users chose to interact more with contents compressed at high bitrates when compared to low bitrates.
- We analysed the correlation between MOS scores and average interaction time, separately for perspective and refocused views. We also computed the average interaction

time for each stimulus, regardless of the rendered view. We showed that the average interaction time is strongly correlated with the subjective results, reaching PCC values of 0.96 and SRCC values of 0.87.

This chapter lays the basis of an implicit quality assessment method for light field contents. However, further analysis is needed to prove if the average interaction time can effectively be used instead of explicit quality scores to assess the visual quality of light field contents. Future work can extend the analysis presented in this chapter, incorporate user interaction data in both subjective and objective quality metrics, and integrate visual attention in perceptual compression solutions for light field data.

5 Rendering-dependent quality evaluation for light field contents

In the past chapters, we have presented image-based rendering as a viable method to experience, and thus evaluate, light field contents on traditional 2D screens. However, it is undisputable that there is a need for light field displays on which the data can be natively visualized, fueled by the recent innovations in the realm of acquisition and compression of light field contents.

A promising solution for light field rendering uses a stack of programmable light-attenuating layers in front of a light-emitting source to provide depth cues without the need of glasses [Lanman et al., 2010] [Lanman et al., 2012] [Wetzstein et al., 2012]. As only a few attenuating layers are required to render multiple points of view, the term "compressive display" has been used to define this type of rendering devices. The pattern images to be displayed in each light-attenuating layer can be obtained from the multi-view light field data through Non-negative Tensor Factorization (NTF) [Wetzstein et al., 2012]. Recently, a new method has been proposed to generate the layer patterns from a stack of focused images (focal stack), which greatly reduces the number of images that are needed as input for the tensor displays [Takahashi et al., 2018]. The method was tested in a prototype 3D display to prove its efficacy [Kobayashi et al., 2017].

Testing the visual quality of compressed and uncompressed light field contents on native light field display is of extreme importance in future development of both new rendering methods, as well as new compression solutions. However, the limited availability of light field displays hinders the assessment of their visual quality. Moreover, hardware limitations in prototype models considerably lessen the perceptual QoE in consuming light field contents. Being able to simulate light field multi-layer rendering in a virtual environment is thus helpful in conducting evaluation of visual quality for light field displays in an ideal scenario.

We have recently proposed a framework to conduct quality assessment of light field contents rendered through a tensor display simulator in 2D screens [Viola et al., 2019]. Through a Graphical User Interface (GUI) the layer patterns composing the multi-layer tensor displays are simulated in a 3D environment. By interacting with the mouse, users can experience the

(a) Bikes (b) Danger de mort (c) Stone Pillars Outside (d) Fountain & Vincent (e) Friends

Figure 5.1 – Central perspective view from each content used in the test.

light field from different points of view. A more detailed description of the framework can be found in Annex C. The framework supports the use of different single- and double-stimulus subjective methodologies; for the latter type, it can be chosen to display the stimuli in a side-by-side fashion, or to visualize them intermittently in the same position on the screen through the use of the keyboard. The second variant is particularly useful when the size of the 2D screen does not support the visualization of both stimuli in full resolution, thus avoiding to resort to cropping. It also allows to assess the impairment of the stimuli for nearly-lossless compression schemes, similarly to the evaluation procedures detailed in ISO/IEC 29170-2 (AIC Part-2) Draft Amendment 2 [ISO/IEC 29170-2, 2015]. However, for higher compression ratios it may be preferable to use the side-by-side variant, as it allows to visualize both stimuli at a glance, as well as granting a higher level of perceptual masking.

In this chapter we analyse the effect of using different double-stimulus variants to perform subjective quality assessment of light field contents, rendered through a simulated compressive display. We also perform the comparison between the results obtained with the simulator and a prototype multi-layer display. To account for cross-cultural variance among different testing groups, we perform the same experiment in two different laboratory settings, and we analyse the correlation among the scores. For our test, we use three compression approaches modeled on the unique opportunities given by multi-layer displays. An in-depth analysis of the compression approaches and their coding efficiency will be given in Chapter 8.

5.1 Experiment design

In this section we will describe the subjective quality experiment we conducted to perform the analysis on quality assessment for simulated multi-layer displays. In particular, we first list the dataset and the coding conditions. We then describe the lab settings in which the tests were conducted, as well as the employed methodologies. Finally, we give an overview of the statistical analysis we conducted on the gathered data.

5.1.1 Dataset and coding conditions

Five light field contents were selected from a publicly available database [Řeřábek and Ebrahimi, 2016]. The contents were acquired with a Lytro Illum camera and processed using the Light Field Matlab Toolbox [Dansereau et al., 2013][Dansereau et al., 2015] to obtain a stack of 15×15 perspective image, each having a resolution of 625×434 pixels. Color and gamma corrections were applied on each perspective image for the rendering. To avoid unwanted distortions caused by the lenslet structure of the Lytro Illum camera, only the 9×9 central perspective views were selected for the test. The central perspective view from each content is displayed in Figure 5.1.

Considering the peculiarities of our rendering system, three viable alternatives for light field compression were employed. The first arranges the 9×9 perspective views in a pseudo-temporal video sequence, which is subsequently encoded. The layer patterns needed for rendering are then created at the receiver side, after decoding the compressed views. The second method creates the layer patterns at the encoder side; such layer patterns are arranged in a pseudo-temporal video sequence and compressed. At the receiver side, the decoded layer patterns can be directly rendered without ulterior processing. Finally, the third solution creates a focal stack of refocused images from the perspective views. The focal stack is then arranged in a pseudo-temporal sequence, compressed and transmitted. At the receiver side, the layer patterns are created from the focal stack. For all three solutions the state-of-the-art video encoding standard HEVC was employed for the compression, to ensure a fair comparison. More information can be found in Chapter 8.

The layer patterns were created using the software implementation presented in [Takahashi, 2018]. To create the focal stack, the Light Field Matlab Toolbox was employed [Dansereau et al., 2013][Dansereau et al., 2015]. In our validating test, the number of layers was fixed to $L = 3$.

The compression solutions were evaluated at four bit-rates, namely $R1 = 537$ kB, $R2 = 134$ kB, $R3 = 67$ kB, and $R4 = 27$ kB, corresponding to 0.2, 0.05, 0.025 and 0.01 bpp, respectively. The bpp are computed with respect to the original size of the 9×9 perspective views. The bit-rates were carefully chosen to cover the visual quality space while providing reasonable and fair comparison among the listed compression solutions.

5.1.2 Subjective methodologies

For our experiments, the DSIS with 5-point grading scale (*5-Imperceptible, 4-Perceptible but not annoying, 3-Slightly annoying, 2-Annoying, 1-Very annoying*) was selected, according to the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. We tested two variants of the same methodology. In Variant A, a side-by-side presentation of the stimulus under test along with the uncompressed reference is employed. Subject can visualize both contents at a glance and give their rating based on the perceived impairment. The assessment depends on

observing the same region of interest through eye and head movement to be able to detect the impairments of the test content with respect to the reference; thus, small artifacts, such as different noise distributions, may be masked while using this variant. The variant is especially useful when the solution under test does not rely on pixel-based accuracy, but on perceptual models (see [Verhack et al., 2017]).

Variant B presents a user-driven intermittent presentation of the impaired and reference stimuli. This variant allows to compare the same region of interest in both stimuli without head or eye movements, by switching between the two contents; as such, it is particularly suitable for immersive multimedia in which a split screen would create unnatural effects and head movements could warrant unwanted consequences, such as omnidirectional imaging [Perrin et al., 2017]. It is also to be preferred when testing just noticeable differences which would not be captured by a side-by-side presentation.

For both variants, participants were asked to rate the quality of the test stimuli when compared to the uncompressed reference. For Variant A, they were informed beforehand on which side of the screen the reference would be displayed, and its position on the screen was fixed for the duration of the test. For Variant B, participants could access the reference content by pressing a specific key, and they could return to the test content by pressing another designated key. Participants were only allowed to give a score when the test contents was being rendered on the screen, and at least one full switch between test and reference stimulus was required to perform the rating. In order to accustom the participants with what distortions to expect in the test images, a training session was organized before the experiment. Three training samples, created by compressing one additional content on the test bit-rates, were manually selected by expert viewers.

All the compressed stimuli were shown in one session. Additionally, two types of hidden reference per content were added to the test: one consisted in the layer patterns generated from the uncompressed stack of perspective views (which was also used as the explicit reference), while the other was created from the uncompressed focal stack. The hidden references were added to account for the artifacts derived from the chosen layer generation method. A total of 70 stimuli were evaluated in each session. The display order of the stimuli was randomized for each participant, and the same content was never displayed twice in a row.

5.1.3 Test environments

Two laboratory settings were used for our tests, in the facilities of the École Polytechnique Fédérale de Lausanne (EPFL) and Nagoya University (NU).

In EPFL, a laboratory for subjective video quality assessment, which was set up according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012], was used for the test. A 27-inch Apple Display with native resolution of 2560 × 1440 pixels was used. The monitor settings were adjusted according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2

Table 5.1 – Test environments and specifications.

Environment	University	Display	No. contents	No. codecs	No. bitrates	No. subjects	Methodology	Approach
Controlled lab setting	EPFL	Apple display	5	3	4	20	DSIS	Side by side
Controlled lab setting	EPFL	Prototype	5	3	4	11	DSIS	Intermittent
Semi-controlled lab setting	NU	Apple display	5	3	4	17	DSIS	Side by side
Semi-controlled lab setting	NU	Prototype	5	3	4	11	DSIS	Intermittent

brightness, and minimum black level of 0.2 cd/m^2. The controlled lighting system in the room consisted of adjustable neon lamps with 6500 K color temperature against mid-grey background walls. The illumination level measured on the screens was 18 lux. Conforming to requirements in ITU-R Recommendation BT.2022 [ITU-R BT.2022, 2012], the distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content. However, subjects were allowed to move further or get closer to the screen. Both Variants A and B were tested in the EPFL facilities, in two separate tests. The order of the test was randomized for each participants, to minimize the influence of employing one variant before the other on the assigned scores. A total of 20 subjects (10 males and 10 females) participated in the tests, amounting to 20 scores per stimulus per variant. Subjects were between 18 and 35, with a mean age of 23.29 years old. Before starting the test, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively.

In NU, a controlled environment was selected to perform the experiment. However, no calibration on the lighting system for the room was conducted. Two displays were used for the tests. First, a 27-inch Apple Display, with the same characteristics of the one employed in EPFL, was used to test DSIS Variant A. A total of 17 subjects (16 males and 1 female) took part in the test. Subjects were between 18 and 35, with a mean age of 24 years old. A prototype multi-layer display was used to perform a pilot evaluation [Kobayashi et al., 2017]. As the resolution of the display did not allow to perform a side-by-side comparison, DSIS Variant B was used for the test. A total of 11 subjects (all males) took part in the test. Subjects were between 18 and 35, with a mean age of 23.28 years old. In both tests, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively. A summary of the specifications for the tests can be found in Table 5.1.

5.2 Statistical analysis

Outlier detection and removal was performed on the results, independently for each test, according to the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. No outlier was detected in either batch of scores. After outlier removal, the MOS was computed for each stimulus, independently for each methodology. The corresponding 95% CIs were computed assuming a Student's t-distribution.

In order to draw a comparison among the different variants, test settings and displays, several fittings were applied to the MOS values, following the ITU-T Recommendation P.1401 [ITU-T P.1401, 2012]. In particular, first order and third order fittings were used to compare the MOS values. RMSE, PCC, SRCC and OR were computed for accuracy, linearity, monotonicity and consistency, respectively.

Multiple comparison tests were performed at a 5% significance level on the raw scores, to determine, for each stimulus, whether the MOS values obtained in different test settings, using different variants or different displays, were significantly different. Furthermore, the percentage of correct estimation, underestimation and overestimation were computed. Additionally, the classification errors were computed using the same multiple comparison test to see if the results obtained with the tested conditions lead, for each pair of stimuli, to the same conclusions [ITU-T J.149, 2004]. In this case, three types of error can be distinguished: false ranking, false differentiation and false tie. False ranking is the most offensive error, and occurs when in the first condition, situation i is better than situation j, whereas in the second condition the opposite is true. False differentiation occurs when in the first condition situation i and j are different, whereas in the second condition they are the same. False tie occurs when with the first condition the two situations are the same, whereas the second condition says they are different.

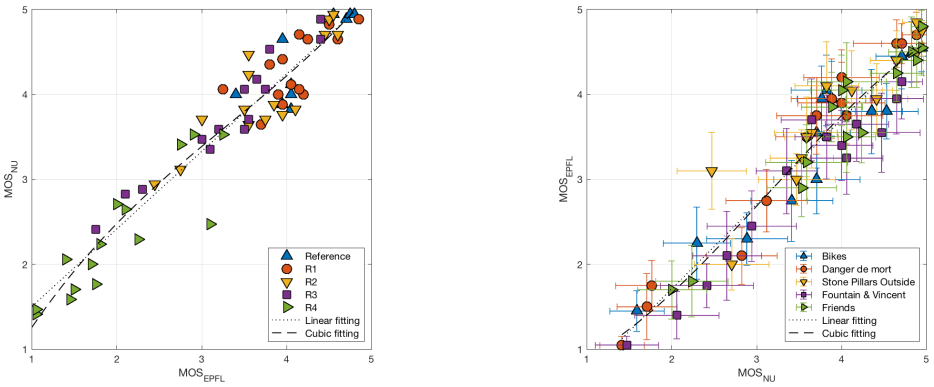
5.3 Results

In this section, results of the comparison between different test conditions are discussed. First, in Section 5.3.1 we compare the Variant A results obtained in two different laboratory settings, to see whether any statistical difference can be caused by different environment settings and cultural backgrounds. Then, the Variant A and B, tested in the same laboratory conditions in EPFL, are compared in Section 5.3.2. Analogously, in Section 5.3.3 we perform a comparison between the results obtained using Variant B in different displays and laboratory settings.

5.3.1 Comparison of different laboratory settings

Figure 5.2 depicts the scatter plot showing the results of the comparison between the MOS scores obtained in the two test settings using DSIS Variant A. In Figure 5.2 (b), the horizontal and vertical bars represent the CIs corresponding to results obtained in NU and EPFL, respectively. To improve visualization, the points are colored based on compression ratio or content. Linear and cubic fittings are shown for both comparisons.

Table 5.2 reports the results of the performance indexes computed on the data. In particular, the performance indexes are computed for every pair of $\widehat{MOS}_X, MOS_Y$, in which X and Y denote the different test settings, and $\widehat{MOS}$ represents the MOS scores obtained after linear and cubic fitting.



(a) MOS_{NU} as function of MOS_{EPFL}.

(b) MOS_{EPFL} as function of MOS_{NU}.

Figure 5.2 – Comparison of MOS values obtained in different laboratory settings, along with linear and cubic fittings. Points are differentiated by compression ratio (a) and by content (b).

Table 5.2 – Performance indexes for the comparison among different laboratory settings.

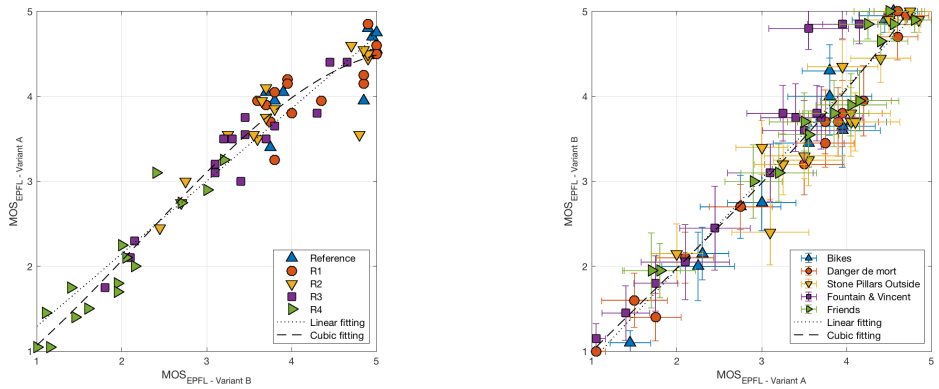
	$[\widehat{MOS}_{EPFL}, MOS_{NU}]$										
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Over Est.**	**Correct Decision**	**False Ranking**	**False Diff.**	**False Tie**
No fitting	0.9542	0.9221	0.4230	4.29%	100%	0.00%	0.00%	88.20%	0.00%	8.41%	3.40%
Linear fitting	0.9542	0.9221	0.2904	1.43%	100%	0.00%	0.00%	88.86%	0.00%	6.25%	4.89%
Cubic fitting	0.9558	0.9221	0.2856	1.43%	100%	0.00%	0.00%	89.15%	0.00%	6.46%	4.39%
	$[\widehat{MOS}_{NU}, MOS_{EPFL}]$										
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Over Est.**	**Correct Decision**	**False Ranking**	**False Diff.**	**False Tie**
No fitting	0.9542	0.9221	0.4230	4.29%	100%	0.00%	0.00%	88.20%	0.00%	3.40%	8.41%
Linear fitting	0.9542	0.9221	0.3035	1.43%	100%	0.00%	0.00%	88.20%	0.00%	3.40%	8.41%
Cubic fitting	0.9551	0.9221	0.3008	1.43%	100%	0.00%	0.00%	87.83%	0.00%	2.36%	9.81%

Table 5.3 – Performance indexes for the comparison among different DSIS variants.

	$[\widehat{MOS}_{EPFL-VariantA}, MOS_{EPFL-VariantB}]$										
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Over Est.**	**Correct Decision**	**False Ranking**	**False Diff.**	**False Tie**
No fitting	0.9578	0.9366	0.3416	10.00%	98.57%	1.43%	0.00%	84.35%	0.00%	4.35%	11.30%
Linear fitting	0.9578	0.9366	0.3268	4.29%	98.57%	1.43%	0.00%	84.35%	0.00%	4.35%	11.30%
Cubic fitting	0.9590	0.9366	0.3220	2.86%	98.57%	1.43%	0.00%	85.34%	0.00%	4.31%	10.35%
	$[\widehat{MOS}_{EPFL-VariantB}, MOS_{EPFL-VariantA}]$										
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Over Est.**	**Correct Decision**	**False Ranking**	**False Diff.**	**False Tie**
No fitting	0.9578	0.9366	0.3416	10.00%	98.57%	0.00%	1.43%	83.11%	0.00%	14.41%	2.48%
Linear fitting	0.9578	0.9366	0.2918	5.71%	100.00%	0.00%	0.00%	84.89%	0.00%	9.73%	5.38%
Cubic fitting	0.9680	0.9366	0.2545	4.29%	100.00%	0.00%	0.00%	89.19%	0.00%	3.27%	7.54%

Table 5.4 – Performance indexes for the comparison among the multi-layer display and the simulator.

	$[\widehat{MOS}_{Simulator}, MOS_{Multi-layerdisplay}]$										
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Over Est.**	**Correct Decision**	**False Ranking**	**False Diff.**	**False Tie**
No fitting	0.5244	0.5817	1.1277	44.29%	72.86%	20.00%	7.14%	52.46%	3.73%	27.25%	16.56%
Linear fitting	0.5244	0.5817	0.9734	48.57%	61.43%	15.71%	22.86%	47.08%	3.11%	17.27%	32.55%
Cubic fitting	0.6008	0.5906	0.9139	42.86%	68.57%	15.71%	15.71%	55.78%	0.62%	18.47%	25.13%
	$[\widehat{MOS}_{Multi-layerdisplay}, MOS_{Simulator}]$										
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Over Est.**	**Correct Decision**	**False Ranking**	**False Diff.**	**False Tie**
No fitting	0.5244	0.5817	1.1277	44.29%	72.86%	7.14%	20.00%	52.46%	3.73%	16.56%	27.25%
Linear fitting	0.5244	0.5817	0.9679	50.00%	74.29%	4.29%	21.23%	46.71%	1.12%	4.60%	47.58%
Cubic fitting	0.6270	0.6023	0.8855	42.86%	77.14%	4.29%	18.57%	56.60%	1.53%	11.59%	30.27%



(a) $MOS_{EPFL-VariantA}$ as function of $MOS_{EPFL-VariantB}$.

(b) $MOS_{EPFL-VariantB}$ as function of $MOS_{EPFL-VariantA}$.

Figure 5.3 – Comparison of MOS values obtained with different methodologies, along with linear and cubic fittings. Points are differentiated by compression ratio (a) and by content (b).

As clearly shown in Figure 5.2, the scores obtained in the two test settings are strongly correlated. In particular, linear regression performed on the $[\widehat{MOS}_{NU}, MOS_{EPFL}]$ pair reports a slope of 0.9974 and an intercept of -0.2830, which indicates that while a strong correlation can be seen between the scores obtained in the two test settings, ratings obtained in EPFL are consistently lower than their NU counterpart. Results of the performance indexes computed on the MOS pairs confirm the strong correlation between the scores obtained in the two test settings. In particular, cubic regression seems to give the best results among the fittings applied to the MOS pairs, with a PCC index of 0.9558 and 0.9551 for the $[\widehat{MOS}_{EPFL}, MOS_{NU}]$ and $[\widehat{MOS}_{NU}, MOS_{EPFL}]$ pair, respectively. The high values of SRCC show a strong monotonicity trend between the two sets of ratings, which confirms the strong correlation proven by the low RMSE and OR values. Finally, the multiple comparison tests show that, although in the two lab settings statistically equivalent scores are given for the same stimulus, as proven by the Correct Estimation index being always at 100%, scores given in the EPFL lab setting tend to be more descriminative, leading to reporting a false tie (i.e., they were considered statistically different in the EPFL case) for 8.41% of stimuli that were deemed statistically equivalent in the NU laboratory setting. However, the effect can be explained by the fact that in the first case, a larger number of people was used to perform the test. Thus, larger CIs are obtained.

The strong correlation among the scores obtained using the same variant across two different laboratory settings shows that the difference in test environment does not strongly determining the distribution of the scores. However, the presence of a certain bias among the scores indicates that further tests are needed to assess the impact of cross-cultural differences, as well as different test environments, in the subjective assessment of light field contents.

5.3.2 Comparison of different DSIS variants

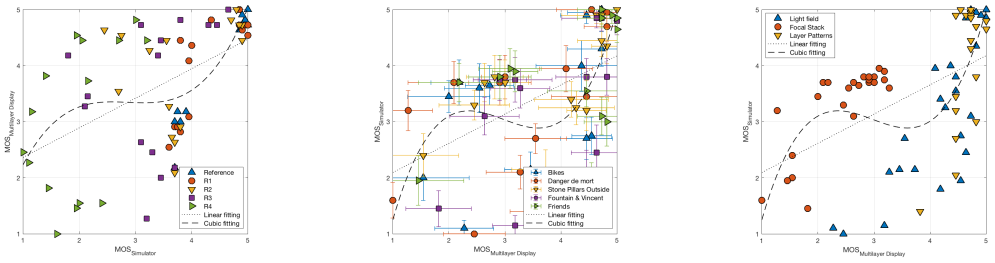
Figure 5.3 shows the results of the comparison between Variant A and B of the DSIS test performed in the EPFL laboratory setting, with linear and cubic fittings. A strong linear correlation can be observed for the scores obtained with the two variants. In particular, for low values of MOS the points are laying on the $y = x$ line, proving that very similar scores are given for heavily compressed contents. However, it can be clearly observed that for lower compression ratios, several contents which were scored on a range from 3 to 5 in the DSIS Variant A, had the same score in Variant B (see Figure 5.3 (a), for MOS values of 5 in the x-axis). This indicates that some artifacts were perceived while the contents were shown side by side, perhaps mistakenly, as switching between reference and test content showed no difference among the two. The CIs among larger scores also seem to be sensibly smaller for Variant B with respect to Variant A (see Figure 5.3 (b), for MOS values of 5 in the y-axis).

Table 5.3 shows the results of the performance indexes for the sets of scores obtained with both variants. Results of PCC and SRCC confirm the strong correlation among the two sets of scores. However, it is interesting to see that the two variants are not always in perfect agreement, as shown by the value of Correct Estimation = 98.57% for all fittings applied to the pair $[\widehat{MOS}_{EPFL-VariantA}, MOS_{EPFL-VariantB}]$. In particular, Variant A seems to lead to significantly lower scores than Variant B for a small percentage of cases, in accordance to what has been seen in Figure 5.3. Results from the multiple comparison show that, while the sets of scores show a significant amount of agreement (more than 84% of time), Variant B leads in general to more differentiation among pairs of scores (10% differentiation over Variant A, versus a 4.3% differentiation of Variant A over Variant B for $[\widehat{MOS}_{EPFL-VariantA}, MOS_{EPFL-VariantB}]$). The trend is only overturnt when considering the cubic fitting for pairs $[\widehat{MOS}_{EPFL-VariantB}, MOS_{EPFL-VariantA}]$, for which Variant A is more discriminative than Variant B (7.54% differentiation over Variant B, versus 3.27% differentiation of Variant B over Variant A). False ranking, the most offensive error, is never encountered.

Results show that, while the two variants are strongly correlated and give agreeable scores for the majority of cases, they do lead to slightly different variations of scores. In particular, Variant A leads to underestimation in a small percentage of cases, and is in general less discriminative. This is very likely to be associated with high MOS scores: for contents that are compressed at nearly transparent quality levels, Variant A leads to more confusion in the assignment of the scores, as shown in Figure 5.3; for those compression ratios, employing Variant B may be the preferrable choice.

5.3.3 Comparison of different displays

Figure 5.4 shows the results of the comparison between the DSIS Variant B tests, as performed in the EPFL laboratory setting using the simulator, and in the NU laboratory using a prototype



(a) $MOS_{Simulator}$ as function of $MOS_{Multi-layerdisplay}$.

(b) $MOS_{Multi-layerdisplay}$ as function of $MOS_{Simulator}$.

(c) $MOS_{Multi-layerdisplay}$ as function of $MOS_{Simulator}$.

Figure 5.4 – Comparison of MOS values obtained with different displays, along with linear and cubic fittings. Points are differentiated by compression ratio (a), by content (b), and by compression solution (c).

multi-layer display. Points are divided by compression ratio, content and compression solution, with linear and cubic fittings. Table 5.4 shows the results of the performance indexes for the sets of scores obtained with both displays.

Results of the comparison show that poor correlation is achieved between the results obtained using the simulator and the results associated to the multi-layer display, with PCC values as low as 0.5244. By observing Figure 5.4 (a) and (b), no visible trend can be observed regarding the compression ratios or contents employed in the tests; rather, it is easy to notice a clear pattern regarding the way scores are distributed in relation to the compression solution that was employed. In particular, when the multi-layer display was employed for the test, no difference seems to be perceived in the quality of compressed layer patterns, regardless of their compression ratio (Figure 5.4 (c), yellow points); in fact, the large majority of the points lay between MOS values of 4 and 5, whereas they span the entire axis when the simulator is used. A similar trend can be observed for part of the contents that employed traditional light field compression, although in some cases, values of MOS close to 2 were given, when the contents were compressed at the highest compression ratio (compare with Figure 5.4 (a)). It is interesting to see that, when the focal stack method was employed to compress the contents, the MOS scores seldom reached values higher than 3 when the multi-layer display was used. However, the method did not reach transparent quality either when the simulator was selected for the test: indeed, the MOS values always stop short of 4. These results will be commented in more detail in Chapter 8.

Multiple comparison results show that the scores obtained in the two tests were statistically equivalent in 60 – 75% of the cases, depending on what fitting has been applied to the scores. More importantly, the two tests seem to agree on the ranking to be given to the various stimuli on around 50% of the cases. The results are most likely caused by the hardware limitations of the prototype display, which do not allow to differentiate among compression artifacts when the layer patterns are directly compressed. In general, it appears that when the multi-layer

display is used, the choice of generating layer patterns from either the light field data or the focal stack has a greater impact on the visual quality, than the compression ratio chosen to encode the data. Using the simulator, on the other hand, allows to more easily differentiate among different levels of compression. This is likely due to the fact that the simulator offers an ideal scenario for multi-layer rendering; the same cannot be said for the prototype display, whose LCD panels did not achieve the same level of transparency, thus affecting the quality of the rendered contents.

5.4 Conclusions

In this chapter we presented a thorough comparison of different quality assessment scenarios for subjective evaluation of light field contents on multi-layer tensor displays. We performed different sets of experiments in two separate laboratory settings, using both a prototype tensor display and a simulator for 2D screens.

Our contributions are the following:

- We propose two variants for the DSIS subjective test methodology. One, called Variant A, presents the two stimuli side-by-side, whereas the other, called Variant B, offers to alternatively see the two stimuli by switching between them.
- We perform multiple tests in two different laboratory settings to assess different test conditions. We start by analyzing if cross-cultural differences, as well as different light and environment conditions, can affect the results, using Variant A of the DSIS methodology. We show that the two sets of scores are highly correlated, although scores collected in one facility are more positively biased with respect to the other laboratory setting.
- We then study the correlation among the two DSIS variants, performing two separate tests in the same laboratory setting. Results show that, while the scores associated to the two tests are strongly correlated, more uncertainty is associated with near-transparent quality when using Variant A with respect to Variant B.
- We finally perform a comparison between scores obtained using a simulator and the ones obtained through the use of a prototype multi-layer displays. Due to hardware limitations, only Variant B is assessed. It is shown that poor correlation is achieved between the two sets of scores. In-depth analysis of the results indicates that, when using the prototype display, subjects are less sensitive to compression artifacts, when compared to the use of a simulator. In particular, it appears that when the multi-layer display is used, the method employed to generate the layer patterns has a greater impact on the scores, with respect to the compression ratio.

Further work is needed to confirm whether a different DSIS variant could lead to different results in the comparison between displays. Moreover, it is worth analysing how the visual

quality of light field rendering is affected by the hardware limitations of the available multi-layer displays, for example, by doing a benchmarking of existing displays. Regarding the topic of which variant of the DSIS methodology should be used to perform subjective quality assessment of light field contents, the choice should be informed by the type of artifact under assessment. For near-lossless levels of distortion, Variant B seems to be the preferrable choice, whereas for more evident levels of distortion both variants can be successfully employed.

Part II

Comparison and evaluation of compression solutions for light field contents

6 Evaluation of state-of-the-art compression solutions for light field coding

Disclaimer: Some of the contents of this chapter were adapted from the following articles, with permission from all co-authors and publishing entities:

Viola, Irene, Martin Řeřábek, Tim Bruylants, Peter Schelkens, Fernando Pereira, and Touradj Ebrahimi. "Objective and subjective evaluation of light field image compression algorithms." In Picture Coding Symposium (PCS), 2016, pp. 1-5. ©2016 IEEE.

Viola, Irene, and Touradj Ebrahimi. "Quality assessment of compression solutions for light field image coding." In 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW). ©2018 IEEE.

Personal contribution: The subjective assessment tests were designed with the help of my co-authors and the experts in the JPEG community. I performed the experiments and curated the analysis.

Finding new solutions to tackle the problem of perceptually efficient light field compression has been an ongoing effort for numerous years. Countless algorithms have been proposed to reduce the amount of data required to store and transmit light field contents, while holding under consideration the importance of maintaing a good visual quality. Such algorithms are usually tested through subjective or, more often, objective means, in order to document the improvement with respect to the state of the art. We have given a general overview of related works in light field compression in Chapter 2; for how exhaustive our inquiry was, we are sure many other solutions were not mentioned, and many others will continue to be proposed.

Frequently, comparing encoding approaches is made arduous by the variety of coding conditions (not to mention raw data) on which each of them is tested. Benchmarking upcoming compression algorithms against the state of the art thus becomes a nearly impossible feat. Even more so is discerning which solution works best under an assortment of conditions that spans the perceptual space, on contents that faithfully represent the challenges the encoding algorithms need to face.

In recent years, two call for proposals have been issued to collect compression solutions for

light field contents under the framework of grand challenges. The goal of grand challenges is to collect the best solutions available at the time of the declaration and compare them in a fair, reliable and reproducible way to determine the most efficient approach to solve a predefined problem, according to predefined criteria. In the case of the ICME 2016 Grand Challenge on Light Field Compression and the ICIP 2017 Grand Challenge on Light Field Coding, the most important criterion was deemed to be perceptual quality, as assessed by both objective quality metrics and subjective evaluations. A set of conditions were defined for the participants to ensure a fair comparison on equal grounds. All the proposed solutions were assessed under rigid conditions of anonymity to avoid biases and partisanship.

In this chapter, we present the outcomes of the aforementioned grand challenges for compression of light field contents. Results are useful as both a snapshot of the state of the art at a given moment, and as a survey on how various solutions may perform differently based on the coding conditions.

6.1 ICME 2016 Grand Challenge

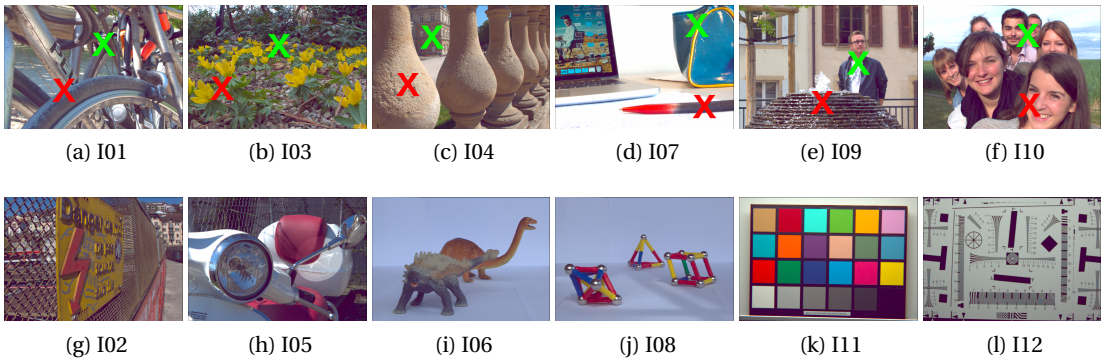
The ICME 2016 Grand Challenge was issued in January 2016 to collect new compression solutions for lenslet-based light field images, and to evaluate them using both objective and subjective quality assessment methodologies [ISO/IEC JTC 1/SC29/WG1 JPEG, 2016]. The grand challenge was focused on compression schemes for raw light field images acquired with a lenslet-based plenoptic camera, specifically, a Lytro Illum plenoptic camera. Data preparation and coding conditions were briefly introduced in Chapter 3.1.1; we report it again here for completeness.

6.1.1 Dataset and coding conditions

As input for the grand challenge, light field images created with a Lytro Illum plenoptic camera were selected. In particular, proponents were asked to compress a lenslet image, which was created from the raw 10-bit sensor data by applying devignetting, demosaicing, clipping to 8-bit and color space conversion from RGB444 to YUV420. The challenge required submission of compression and decompression algorithm capable of processing the given image data according to the end-to-end chain depicted in Fig. 6.2. Specifically, the proponents were asked to implement steps from A to A'.

Twelve lenslet light field images from a publicly available dataset [Řeřábek and Ebrahimi, 2016] were selected for the grand challenge. The central perspective view of each content is depicted in Figure 6.1.

For the objective and subjective evaluations, the decompressed lenslet image was converted to a stack of all-in-focus perspective views (light field data structure) using the Matlab implementation of the Light Field Toolbox v0.4 [Dansereau et al., 2013][Dansereau et al., 2015]. Each



(a) I01 (b) I03 (c) I04 (d) I07 (e) I09 (f) I10

(g) I02 (h) I05 (i) I06 (j) I08 (k) I11 (l) I12

Figure 6.1 – Central all-in-focus view from each content used in the experiments. Refocused points marked in green (slope 1) and red (slope 2). ©2016 IEEE

perspective view was created by selecting and aligning samples from each micro-lens element that supported a particular point of view. The resulting light field data structure is a 5-D array with dimensions of $15 \times 15 \times 434 \times 625 \times 3$, in which 15×15 is the number of perspective views, 434×625 is the resolution of each view, and 3 corresponds to the color channels. Color and gamma corrections were applied to each perspective view. The same pipeline was employed to generate the reference light field data structure from the uncompressed YUV420 lenslet image.

The performance of the proposed compression algorithms was evaluated at four fixed compression ratios, namely $R1 = 10:1$ (1 bpp), $R2 = 20:1$ (0.5 bpp), $R3 = 40:1$ (0.25 bpp), $R4 = 100:1$ (0.1 bpp). The ratios were computed with respect to the size of the raw data obtained from the camera.

Overall, seven submissions were received as responses to the call for proposals in the framework of the grand challenge. Only five of them were accepted in the reviewing process for further evaluation. Proponents were assigned a random number (P1 to P5) to anonymize their identity. In general, two main coding approaches were proposed. The first approach uses a modified version of HEVC Intra encoder to compress the lenslet image by exploiting existent redundancies. The second approach creates the light field data structure prior to coding and then rearranges the sub-aperture images in a pseudo-temporal sequence to be coded with HEVC. In the following paragraphs, we present the submitted algorithms in details. The presentation order does not correspond to the label assigned to each codec.

In [Conti et al., 2016] authors suggested to use HEVC Intra Profile to code the lenslet structure, and to improve its performance by integrating SS compensated prediction and estimation. The proposed solution exploits the correlation between neighbooring micro-images in the lenslet image. The image is partitioned in blocks using HEVC partition patterns. Then, two blocks are selected for predicting the current block, one given by best block matching in the search window and the other selected by searching for best linear combination between the first selected block and a second block in the same window. The best among the two is selected

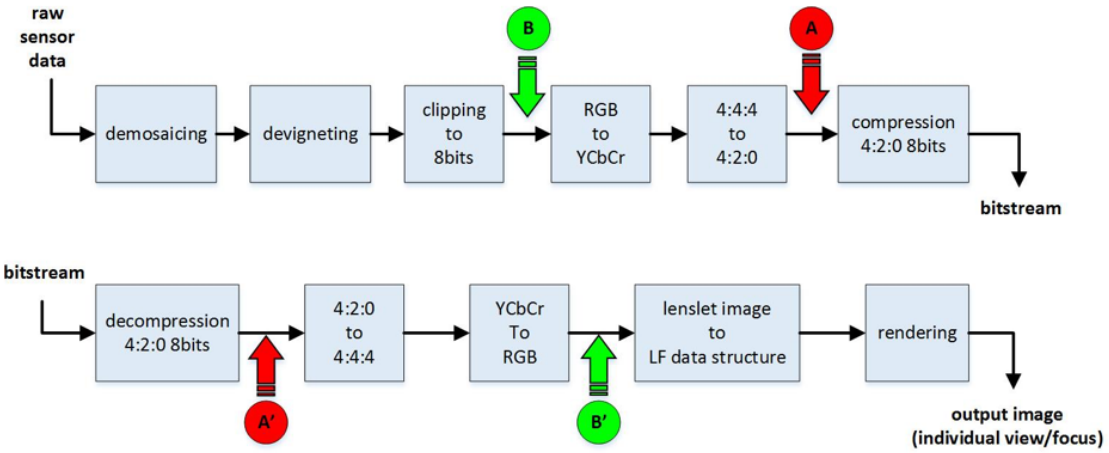


Figure 6.2 – End-to-end chain for compression and decompression of light field lenslet image.

for SS estimation.

The same approach is used in [Monteiro et al., 2016], which integrates SS compensated prediction in HEVC Intra coding, and additionally implements LLE to further improve the compression performance. LLE estimates the current block by solving a least-squares optimization problem to find the best linear combination of k nearest neighbors in a casual search window.

The authors in [Li et al., 2016a] use HEVC Intra profile to encode the lenslet image; however, the conventional intra prediction from reconstructed information is improved by allowing the predictor to use only blocks from its reconstructed neighbors. In addition to that, advanced motion vector prediction is used.

In [Perra and Assuncao, 2016] the chosen approach is to partition the lenslet image into tiles of equal sizes, which are then ordered in a pseudo-temporal sequence using a properly selected scan order. Then the sequence is encoded using HEVC.

Authors in [Liu et al., 2016] use a different approach, and propose a compression of light field images based on pseudo-sequences of sub-aperture images. The lenslet image is first converted from YUV420 to RGB444 color space. Then the lenslet is processed to obtain the multiple views that compose the light field data structure. The views are color and gamma corrected and then converted back to YUV420. A subset of them is then rearranged in a specific coding order that accounts for similarities between adjacent views and coded using the JEM encoder[1].

The proponents were compared to an anchor generated using legacy JPEG, referred to as P0 in the rest of the paper.

[1] https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/tags/HM-16.6-JEM-2.0rc1/

6.1.2 Visual quality assessment

The performance of each proposal was evaluated through both objective and subjective quality assessment. To measure distortions introduced by the compression algorithms, the light field data structure, obtained after compressing and decompressing the lenslet image, is compared to the uncompressed reference, obtained by omitting steps A to A' in Figure 6.2. Both objective and subjective assessments are conducted on reference and test contents after the transformation to light field data structure (output in Figure 6.2).

Objective quality metrics

The metrics chosen to perform the evaluation are PSNR and SSIM, applied separately to individual color channels. The PSNR is computed on the Y channel as follows:

$$PSNR_Y(k,l) = 10\log_{10}\frac{255^2}{MSE(k,l)}, \tag{6.1}$$

in which k and l are the indexes of the sub-aperture images. The $MSE(k,l)$ for each image is computed as follows:

$$MSE(k,l) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j) - R(i,j)]^2, \tag{6.2}$$

where m and n are the dimensions of one sub-aperture image (i.e., $n = 625$, $m = 434$). $I(i,j)$ is the Y value for the selected sub-aperture image in the evaluated light field data structure, whereas $R(i,j)$ is the corresponding value in the reference data structure. In the same way, we can compute the PSNR for the other two channels U and V, obtained after upsampling the color space as depicted in Fig. 6.2. A weighted average [Ohm et al., 2012] is then computed as follows:

$$PSNR_{YUV}(k,l) = \frac{6PSNR_Y(k,l) + PSNR_U(k,l) + PSNR_V(k,l)}{8} \tag{6.3}$$

The mean of sub-aperture images is subsequentially computed to have an average value for PSNR for Y channel and for YUV:

$$\widehat{PSNR}_Y = \frac{1}{(K-2)(L-2)}\sum_{k=2}^{K-1}\sum_{l=2}^{L-1}PSNR_Y(k,l), \tag{6.4}$$

$$\widehat{PSNR}_{YUV} = \frac{1}{(K-2)(L-2)} \sum_{k=2}^{K-1} \sum_{l=2}^{L-1} PSNR_{YUV}(k,l) \tag{6.5}$$

In a similar fashion, the SSIM is computed on the Y channel of each sub-aperture image as follows:

$$SSIM_Y(k,l) = \frac{(2\mu_I\mu_R + c_1)(2\sigma_{IR} + c_2)}{(\mu_I^2 + \mu_R^2 + c_1)(\sigma_I^2 + \sigma_R^2 + c_2)}, \tag{6.6}$$

in which μ_I and μ_R are the average of the Y channel of the two sup-aperture images at index k and l, σ_I^2 and σ_R^2 is the variance, and σ_{IR} is the covariance of the two sub-aperture images in channel Y. $c_1 = (p_1 D)^2$ and $c_2 = (p_2 D)^2$ are two variables to stabilize the division; D is the dynamic range of the pixel values, while $p_1 = 0.01$ and $p_2 = 0.03$ by default.

The SSIM value for the three channels and the mean value is computed following what has already been said for PSNR (equation 6.3, 6.4 and 6.5).

Subjective methodology

For the subjective evaluation, only six contents were chosen, namely, *I01, I03, I04, I07, I09* and *I10*. A thumbnail of the contents is depicted in Figures (a) to (f). The contents were selected by experts among the twelve contents that were used for objective evaluation.

The subjective methodology has been extensively discussed in Chapter 3. A summary of the specifications for the test can be found in Table 3.2.

Subjective Data Processing and Statistical Analysis

Outlier detection and removal was performed on raw scores of naïve subjects according to the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. One subject was found to be an outlier and the corresponding scores were discarded. This led to 17 scores per stimulus. After outlier removal, the MOS was computed for each coding condition j (i.e. for each content, view, proponent and bitrate) as follows:

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} m_{ij}, \tag{6.7}$$

where N is the number of subjects and m_{ij} is the score for stimulus j by subject i.

In order to determine whether the differences between proponents were statistically significant, all the codecs were compared by means of a two-sided Welch's test at 5% significance level, with following hypotheses:

$$H_0 : MOS_{P_A} = MOS_{P_B}$$

$$H_1 : MOS_{P_A} \neq MOS_{P_B},$$

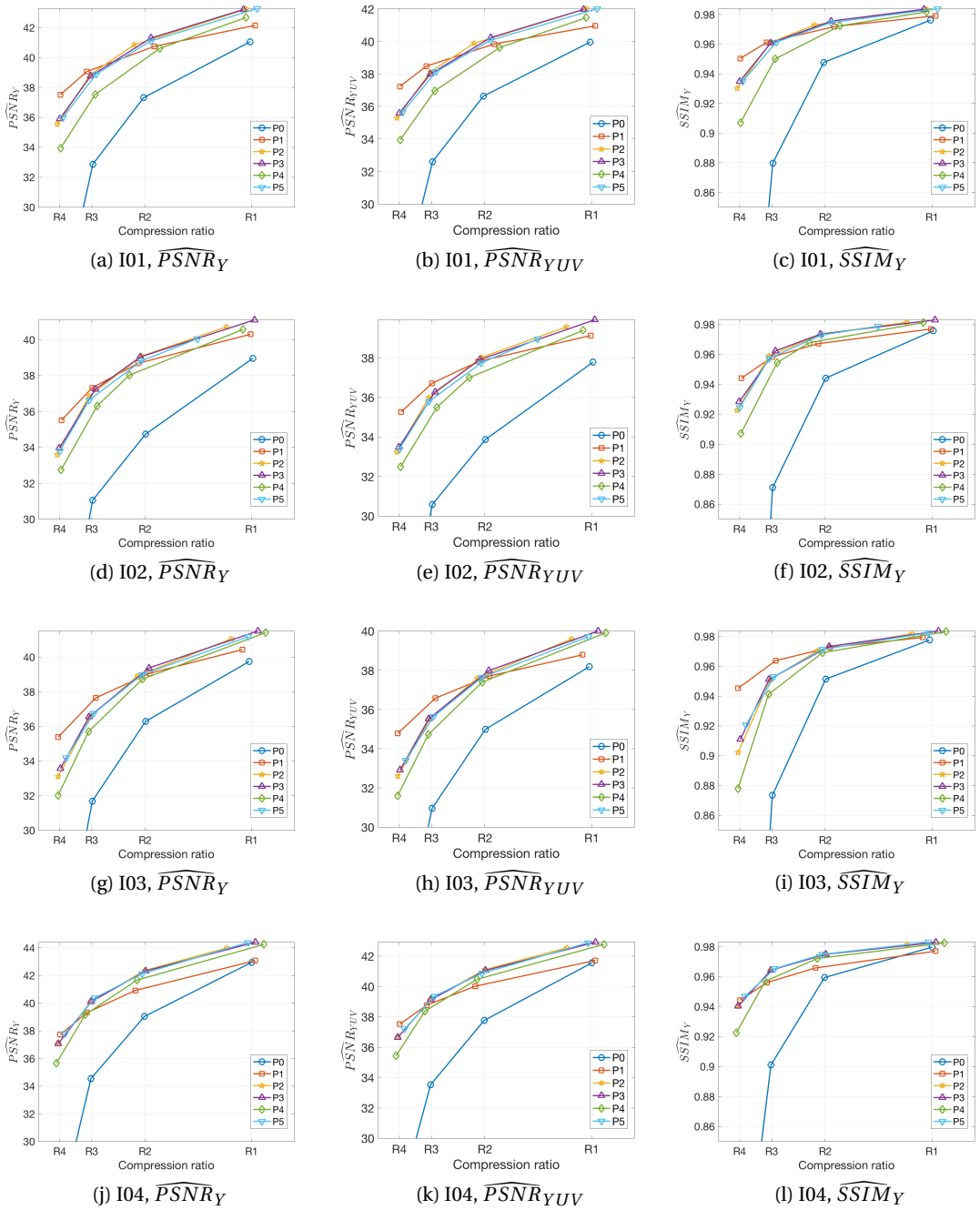
in which P_A and P_B are the proponents that are being compared. If the hypothesis H_0 were to be accepted, it would mean that the difference between means is zero, and that the distribution of difference between mean values follows a t-distribution. On the other hand, if the hypothesis were to be rejected, the conclusion would be that the two values are significantly different. In the test, if the null hypothesis was rejected at 5% significance level, then the two MOS were compared in order to identify which codec performed significantly better. For each content and view, if the hypothesis were to be rejected, the matrix M would be updated as such:

$$M(i,j) = M(i,j) + 1 \text{ if } MOS_i > MOS_j$$

$$M(j,i) = M(j,i) + 1 \text{ if } MOS_i < MOS_j$$

6.1.3 Results

Figures 6.3, 6.4 and 6.5 show the results of the objective evaluation for all contents. The graph has been cropped in order to allow a better analysis of the results. Generally, it can be noticed that for compression ratio $R1$, the difference between the proponents and the anchor is significantly small (around 2 dB for $\widehat{PSNR}_Y$), whereas among the proponents no clear winner can be appointed. It is worth noting that for most contents, P1 largely performs worse at this compression ratio when compared to other proponents, according to metrics $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$, with notable exceptions of contents I05 and I09. When using $\widehat{SSIM}_Y$ as a metric, all proponents and the anchor have similar performance for compression ratio $R1$ for most contents, with notable exceptions for proponent P1. For higher compression ratios the difference among the proponents and the anchor becomes stark. Results clearly show that while for high bitrates JPEG can be considered as a valid alternative to the proposed solutions, for low bitrates it is inadequate for light field compression. For compression ratios $R3$ and $R4$, P1 outperforms the other codecs, gaining around 3 dB for the same bitrate with respect to P4 for metrics $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$. Curves for $\widehat{SSIM}_Y$ show similar results, with P1 outperforming all other proponents for lower bitrates.



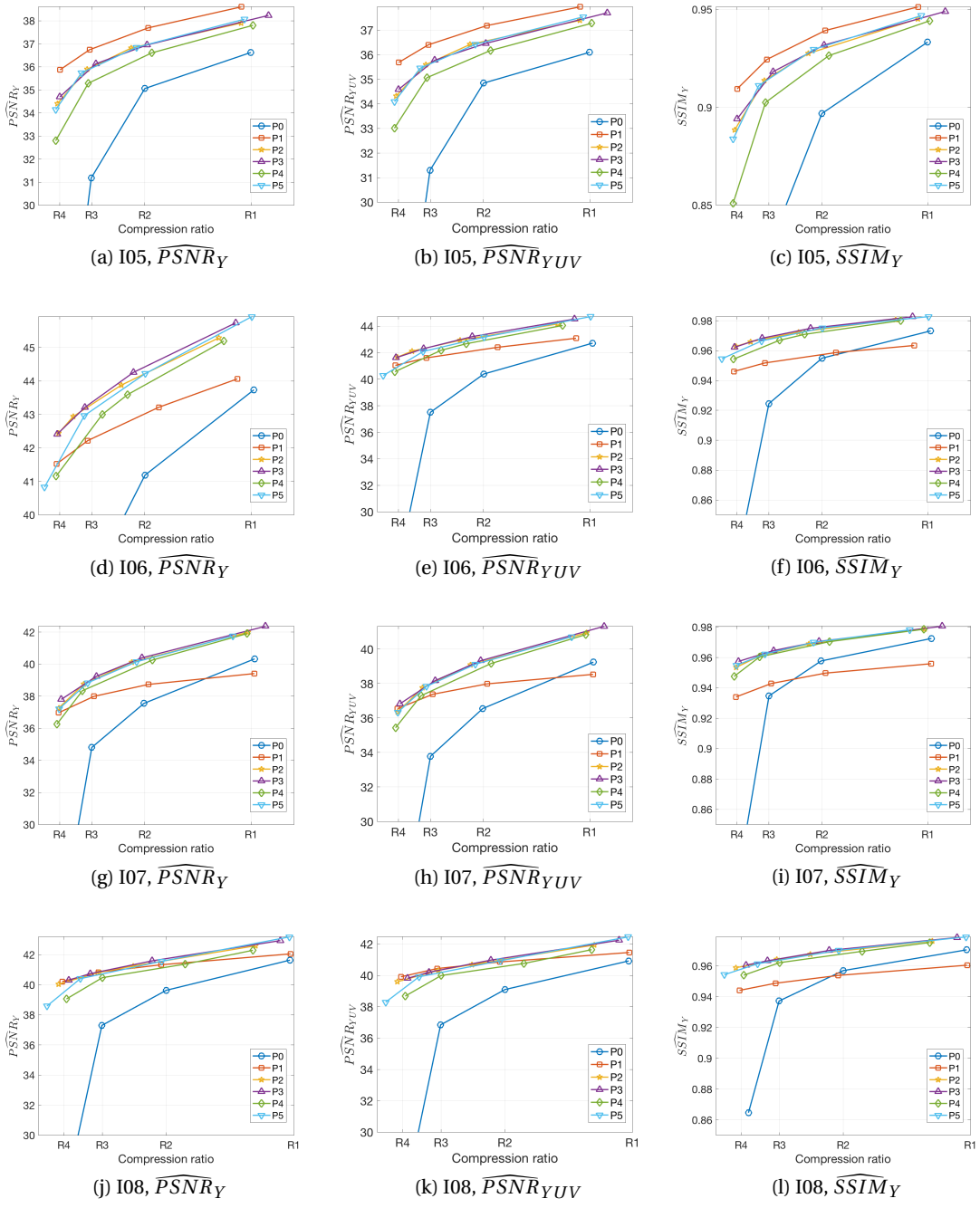
(a) I01, $\widehat{PSNR}_Y$ (b) I01, $\widehat{PSNR}_{YUV}$ (c) I01, $\widehat{SSIM}_Y$

(d) I02, $\widehat{PSNR}_Y$ (e) I02, $\widehat{PSNR}_{YUV}$ (f) I02, $\widehat{SSIM}_Y$

(g) I03, $\widehat{PSNR}_Y$ (h) I03, $\widehat{PSNR}_{YUV}$ (i) I03, $\widehat{SSIM}_Y$

(j) I04, $\widehat{PSNR}_Y$ (k) I04, $\widehat{PSNR}_{YUV}$ (l) I04, $\widehat{SSIM}_Y$

Figure 6.3 – Results of the objective evaluations for contents I01-I04 (rows). $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.



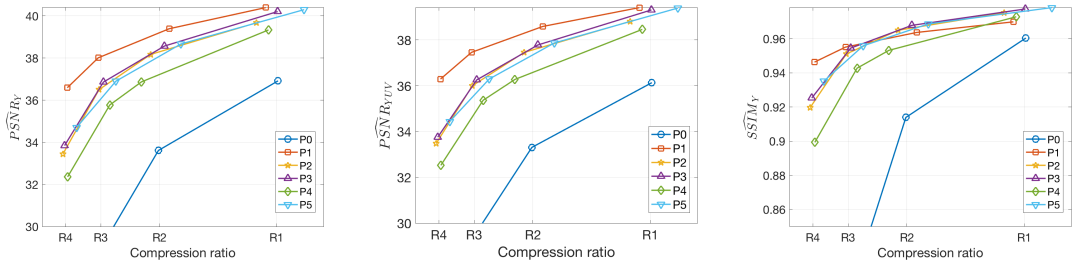
(a) I05, $\widehat{PSNR}_Y$ (b) I05, $\widehat{PSNR}_{YUV}$ (c) I05, $\widehat{SSIM}_Y$

(d) I06, $\widehat{PSNR}_Y$ (e) I06, $\widehat{PSNR}_{YUV}$ (f) I06, $\widehat{SSIM}_Y$

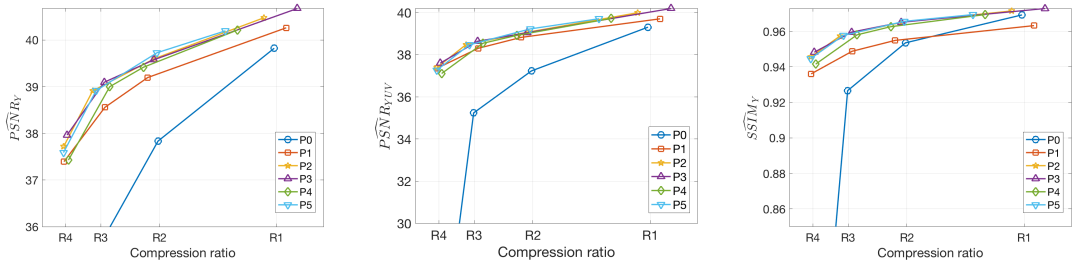
(g) I07, $\widehat{PSNR}_Y$ (h) I07, $\widehat{PSNR}_{YUV}$ (i) I07, $\widehat{SSIM}_Y$

(j) I08, $\widehat{PSNR}_Y$ (k) I08, $\widehat{PSNR}_{YUV}$ (l) I08, $\widehat{SSIM}_Y$

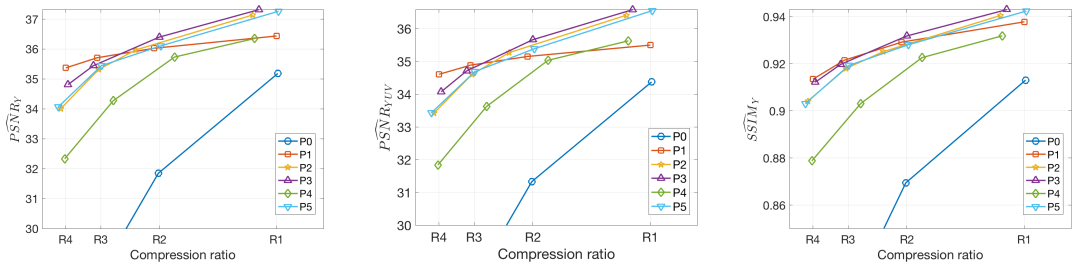
Figure 6.4 – Results of the objective evaluations for contents I05-I08 (rows). $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.



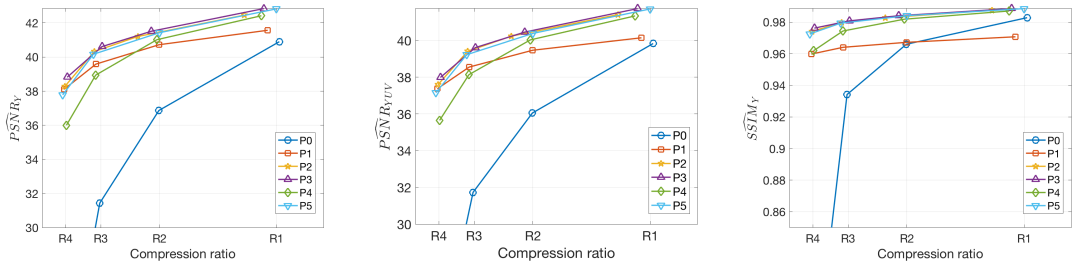
(a) I09, $\widehat{PSNR}_Y$ (b) I09, $\widehat{PSNR}_{YUV}$ (c) I09, $\widehat{SSIM}_Y$



(d) I10, $\widehat{PSNR}_Y$ (e) I10, $\widehat{PSNR}_{YUV}$ (f) I10, $\widehat{SSIM}_Y$



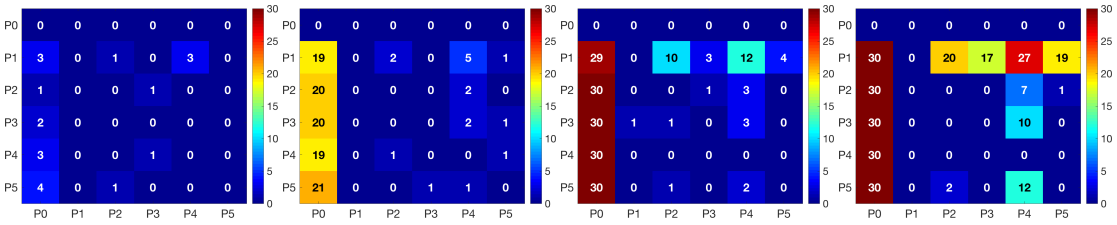
(g) I11, $\widehat{PSNR}_Y$ (h) I11, $\widehat{PSNR}_{YUV}$ (i) I11, $\widehat{SSIM}_Y$



(j) I12, $\widehat{PSNR}_Y$ (k) I12, $\widehat{PSNR}_{YUV}$ (l) I12, $\widehat{SSIM}_Y$

Figure 6.5 – Results of the objective evaluations for contents I09-I12 (rows). $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.

Figures 6.7 and 6.8 depict the results of the subjective evaluation campaign for all views and all contents. The proponents and the anchor are plotted with a full line with respective CIs, whereas the MOS for the uncompressed reference, with corresponding CIs, is shown through



(a) R1 (b) R2 (c) R3 (d) R4

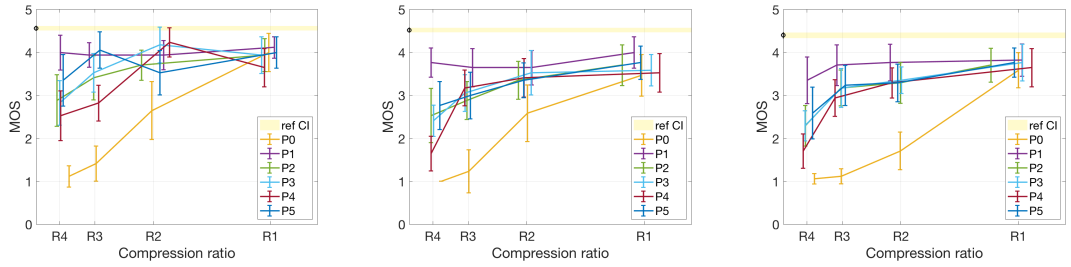
Figure 6.6 – Pairwise comparison of codecs for different bitrates. ©2016 IEEE

a yellow stripe. Figure 6.6 shows for how many contents and views the proponent on the y-axis performs significantly better than the proponent on the x-axis. The minimum value is 0 and the maximum value is 30, corresponding to all possible views and contents.

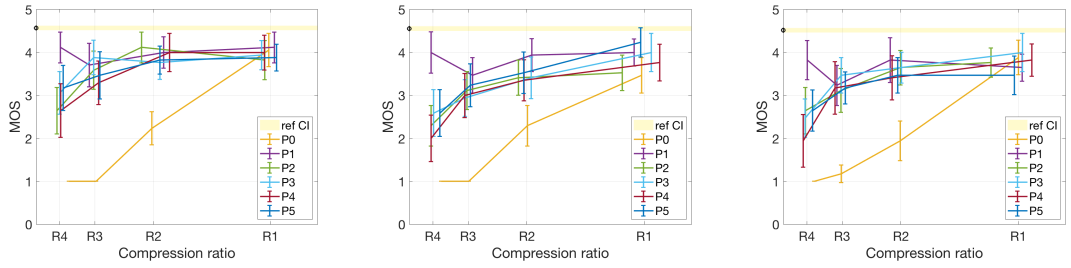
Confirming what was shown by the objective quality metrics, all proponents perform similarly to the anchor for compression ratio $R1$, whereas for lower bitrates they significantly outperform the anchor. Moreover, for high bitrates there is no proponent that performs significantly better than the others (Figures 6.6 (a) and (b)). However, for lower bitrates, similarly to what has been seen for objective results, P1 performs better than other proponents, outperforming them for compression rate $R4$ in more than half of the contents (Figure 6.6 (d)).

As can be seen in Figures 6.7 and 6.8, a significant drop in MOS values can be observed when taking into account perspective views, as opposed to refocused views. The decrease is visible for both, compressed images as well as for uncompressed references. However, the difference of scores between reference and proponents remains constant.

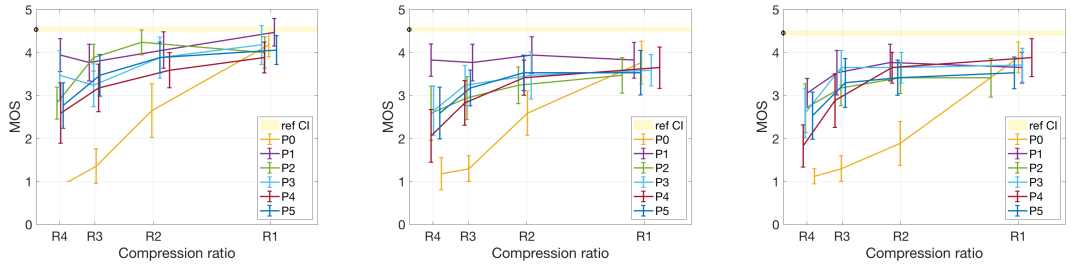
These observations suggest that the viewers found that refocusing the content negatively affects its visual image quality. The topic has been extensively examined in Chapter 3.1.



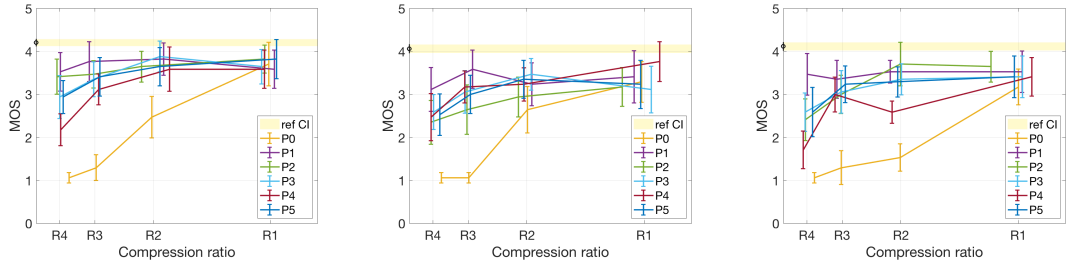
(a) I01, left perspective view (b) I03, left perspective view (c) I04, left perspective view



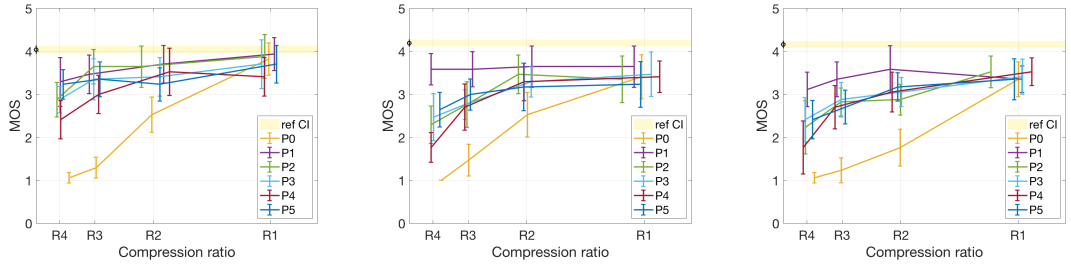
(d) I01, central perspective view (e) I03, central perspective view (f) I04, central perspective view



(g) I01, right perspective view (h) I03, right perspective view (i) I04, right perspective view

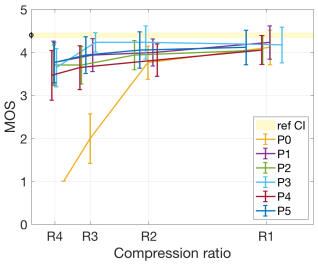


(j) I01, front refocused view (k) I03, front refocused view (l) I04, front refocused view

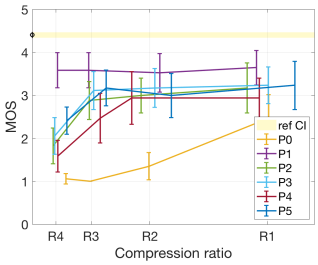


(m) I01, back refocused view (n) I03, back refocused view (o) I04, back refocused view

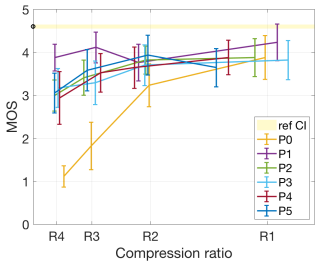
Figure 6.7 – Results of the subjective evaluations for contents I01, I03 and I04 (first, second and third column, respectively). Each row represents the results related to a certain view.



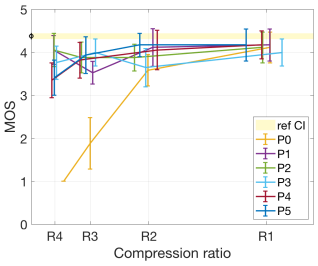
(a) I07, left perspective view



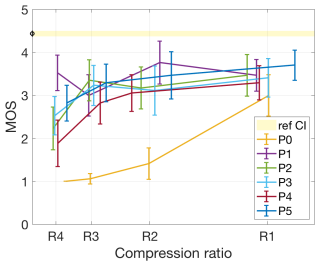
(b) I09, left perspective view



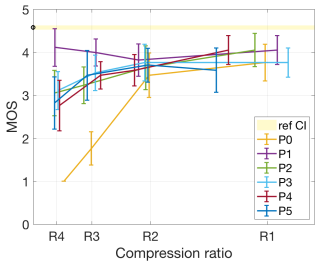
(c) I10, left perspective view



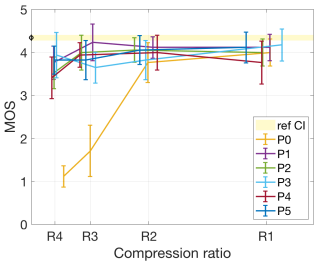
(d) I07, central perspective view



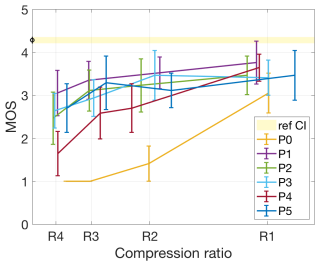
(e) I09, central perspective view



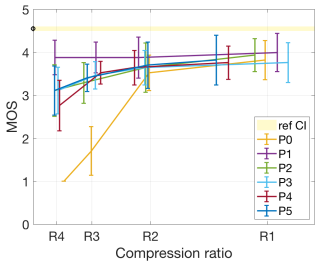
(f) I10, central perspective view



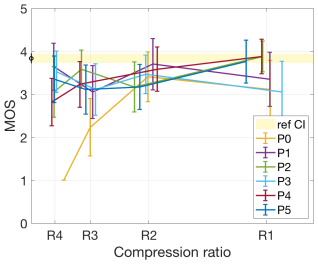
(g) I07, right perspective view



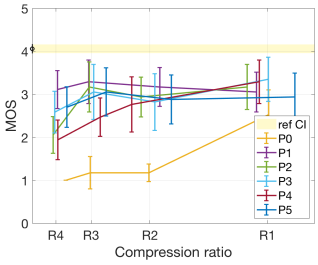
(h) I09, right perspective view



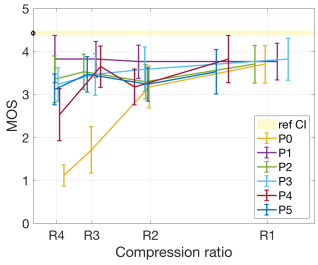
(i) I10, right perspective view



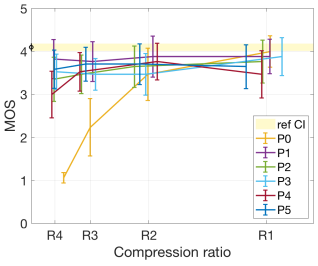
(j) I07, front refocused view



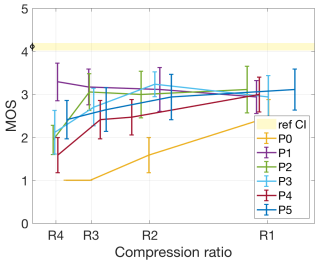
(k) I09, front refocused view



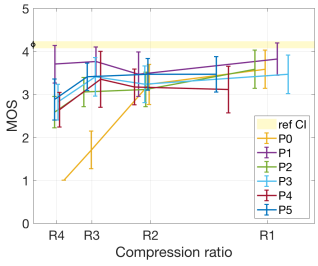
(l) I10, front refocused view



(m) I07, back refocused view



(n) I09, back refocused view



(o) I10, back refocused view

Figure 6.8 – Results of the subjective evaluations for contents I07, I09 and I10 (first, second and third column, respectively). Each row represents the results related to a certain view.

Figure 6.9 – Encoding workflow for lenslet images. ©2018 IEEE

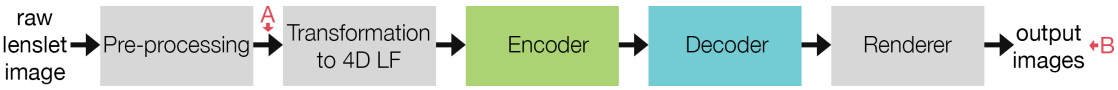


Figure 6.10 – Encoding workflow for perspective views. ©2018 IEEE

6.2 ICIP 2017 Grand Challenge

The JPEG Pleno Call for Proposals in association with ICIP Grand Challenge on Light Field Image Coding was issued in January 2017 to collect new compressing solutions for light field images, and to evaluate them using both objective and subjective quality assessment methodologies. The grand challenge was divided into two main tasks, devoted on compressing light field images acquired with two different technologies, namely a plenoptic (lenslet) camera and a Ultra High Definition (UHD) High Density Camera Array (HDCA) setup. Further information about the requirements for the challenge can be found in [ISO/IEC JTC 1/SC29/WG1 JPEG, 2017].

6.2.1 Dataset and coding conditions

Lenslet camera

For the lenslet-based challenge, proponents were asked to compress light field images acquired with a Lytro Illum plenoptic camera[1], which uses an array of micro-lenses in front of the main sensor. The data obtained from the camera, usually referred to as lenslet image, needs to be processed to be properly rendered, via transformation to a 4D light field structure of perspective views [Levoy, 2006]. For the challenge, the proponents could follow two workflows: one focused on compressing the lenslet image (Figure 6.9), and the other focused on compressing the stack of perspective views obtained after transformation to 4D light field structure (Figure 6.10). Additionally, proponents were asked to provide a renderer, either proprietary or belonging to a third party, that could make the decoded bitstream ready for visualization, supporting their adopted representation model. This step was implemented to collect and assess different representation models for light field rendering.

Five contents were selected from a light field image dataset to be compressed for the grand

[1] https://www.lytro.com/

Table 6.1 – Summary of compression schemes for the lenslet test. ©2018 IEEE

Proponents	Description
HEVC	Anchor: Compression of perspective views using HEVC Main10 (x265 software implementation).
VP9	Anchor: Compression of perspective views using VP9 (reference software).
P01	Compression of perspective views using HEVC and linear approximation prior [Zhao and Chen, 2017].
P02	Compression of perspective views using MV-HEVC [Ahmad et al., 2017].
P03	Compression of lenslet image using JPEG 2000 and depth, disparity and sparse prediction [Tabus et al., 2017].
P04	Compression of perspective views modeled as Gaussian Mixture Model [Verhack et al., 2017].
P05	Compression of lenslet image using optimal arrangement and enhanced illumination model [Jia et al., 2017].

(a) I01 (b) I02 (c) I04 (d) I09 (e) I10

Figure 6.11 – Central perspective view from each content used in the lenslet test. ©2018 IEEE

challenge, namely I01 = *Bikes*, I02 = *Danger_de_Mort*, I04 = *Stone_Pillars_Outside*, I09 = *Fountain_&_Vincent_2* and I10 = *Friends_1* [Řeřábek and Ebrahimi, 2016]. The central view of each content is depicted in Figure 6.11.

Demosaicing and devignetting was applied on the raw camera data to create the 10-bit lenslet images (point A in Figures 6.9 and 6.10). Each lenslet image was then processed using the Light Field MATLAB Toolbox v0.4 [Dansereau et al., 2013, 2015] to create 15×15 10-bit perspective views, which were also color and gamma corrected. Both the lenslet image and the perspective views were given as possible input for the grand challenge. The Light Field MATLAB Toolbox was selected as reference renderer, and the input perspective views constituted the reference B_{Ref}. The performance of the proposed coding algorithms was evaluated on four fixed compression ratios, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, and $R4 = 0.005$ bpp. The ratios were computed with respect to the raw lenslet image size (7728×5368 pixels).

To assess the performance of the proposals, two anchors were created using state-of-the-art video codecs, namely HEVC Main10 and VP9. Following the workflow depicted in Figure 6.10, both codecs perform the compression on the perspective views, which were previously rearranged according to a serpentine order, converted to YCbCr format and downsampled from 4:4:4 to 4:2:2, 10-bit depth. For the first anchor, the HEVC implementation x265 was used[2], while for the second anchor, the VP9 reference software was used to compress the pseudo-temporal sequence[3]. Full description of the command line used to create the anchors can be found in the JPEG Pleno Lenslet Dataset website[4].

[2] https://www.videolan.org/developers/x265.html
[3] https://www.webmproject.org/vp9/
[4] http://grebjpeg.epfl.ch/jpeg_pleno/index_lenslet.html

Overall, a total of six submissions were received as responses for the JPEG Pleno Call for Proposals and the ICIP Grand Challenge. Two of the proposals follow the workflow described in Figure 6.9, whereas four of them adopt the workflow described in Figure 6.10. Additionally, two state-of-the-art video codecs were used as anchors to compare and validate the results. Authors of the first algorithm $P01$ exploit the redundacies in the 4D light field structure of perspective views by estimating a part of them as a weighted sum of other perspective views, adopting a linear approximation prior [Zhao and Chen, 2017]. They use HEVC to encode and transmit part of the views, while non-encoded views are estimated by solving an optimization problem. For algorithm $P02$, authors arrange the perspective views into a multiview structure that can be exploited by the corresponding extension of HEVC, namely MV-HEVC [Ahmad et al., 2017]. They also propose a rate allocation scheme to progressively assign the QPs in order to optimize the performance. Authors of $P03$ design a lenslet-based compression scheme that uses depth, disparity and sparse prediction information to reconstruct the final set of views [Tabus et al., 2017]. The bitrate allocation can be configured to improve the reconstruction by encoding the lenslet image using JPEG 2000, or to allow random access by encoding a subset of views. Authors of $P04$ propose a novel representation of the 4D light field as a multi-modal Gaussian Mixture Model, which can be used to reconstruct the perspective views using only the parameters of the model [Verhack et al., 2017]. For algorithm $P05$, authors propose a lenslet-based encoding scheme that uses a fully reversible transformation to 4D light field to create sub-aperture views, which are then optimally re-arranged and compressed using enhanced illumination compensation in JEM software[5]. Adaptive filtering is then applied to reconstruct the lenslet image [Jia et al., 2017].

The anchors were not evaluated at their maximum reconstruction power, as the reference renderer was used in the workflow ($B_{Max} = B_{Ref}$). Moreover, due to the limitations of their representation model, the authors of $P04$ chose not to submit any results for compression ratio $R1$. Hence, a total of 185 stimuli were used for the evaluation. A summary of the proposals and the anchors can be found in Table 6.1.

High density camera array

In this part of the challenge, proponents were asked to compress a high density array of images acquired by a single camera. Along with providing a compression algorithm that could handle the large image array, proponents were asked to provide a renderer, either proprietary or belonging to a third party, similarly to what was asked in the lenslet part. This step was implemented to collect and assess different representation models for light field rendering.

Four light field contents with high angular density were provided by the Fraunhofer Institute [Fraunhofer Institute, 2017], namely S02 = *TableTop I*, S06 = *TableTop II*, S09 = *Lightfield Production*, and S10 = *Workshop*. Each content was acquired with a Sony Alpha 7 RII robotized

[5] https://jvet.hhi.fraunhofer.de/

Table 6.2 – Summary of compression schemes for the HDCA test.

Proponents	Description
HEVC	Anchor: Compression of perspective views using HEVC Main10 (x265 software implementation).
VP9	Anchor: Compression of perspective views using VP9 (reference software).
P06	Compressed rendering with MR-DIBR [Graziosi et al., 2014].

(a) S02 (b) S06 (c) S09 (d) S10

Figure 6.12 – Example view from each content used in the HDCA test.

camera equipped with a $50mm1.4$ lens, moving with a horizontal step of $4mm$ and a vertical step of $6mm$, covering a plane of $400 \times 120mm$. The precision of each step was $80\mu m$. Each content was comprised of 101×21 perspective views, which were pre-processed to ensure the same lighting conditions and white balance, and aligned to match the acquisition array disposition. One depictive view from each content is shown in Figure 6.12.

Each perspective view was cropped to UHD resolution (3840×2160 pixels), and constituted the reference B_{Ref}. The performance of the proposed coding algorithms was evaluated on four fixed compression ratios, namely $R1 = 0.01$ bpp, $R2 = 0.005$ bpp, $R3 = 0.0025$ bpp, and $R4 = 0.00125$ bpp. The ratios were computed with respect to the size of the entire light field content ($3840 \times 2160 \times 101 \times 21$ pixels).

To assess the performance of the proposals, the same anchors as in the lenslet challenge were selected. Each perspective view forming the light field contents was converted to YCbCr format and downsampled from 4:4:4 to 4:2:2, 10-bit depth. For the first anchor, the HEVC implementation x265 was used[6], while for the second anchor, the VP9 reference software was used to compress the pseudo-temporal sequence[7]. Full description of the command line used to create the anchors can be found in the JPEG Pleno High Density Camera Array Dataset website[8].

Only one full proposal was received as a response for the HDCA part of the JPEG Pleno Call for Proposals. Algorithm *P*06 combines compression and rendering in a unique stage, aptly called Compressed Rendering [Graziosi et al., 2014]. The algorithm takes into account scene geometry and optical properties of the acquiring system to separate the light field data into a texture-plus-depth representation, after performing a visibility test to exclude parts of the

[6] https://www.videolan.org/developers/x265.html
[7] https://www.webmproject.org/vp9/
[8] http://grebjpeg.epfl.ch/jpeg_pleno/index_HDCA.html

image which will not be rendered. The depth map is converted to a disparity map and encoded along with the texture data. On the decoder side, they propose a Multi Reference Depth Image-Based Rendering (MR-DIBR) algorithm to synthetize the perspective views. A summary of the proposal and the anchors can be found in Table 6.2.

6.2.2 Visual quality assessment

All the proposals were assessed through full reference objective quality metrics and subjective evaluations after the rendering stage (point B in Figures 6.9 and 6.10). The reference B_{Ref} was obtained by omitting the encoding and decoding stage in the workflow (shown in green and blue, respectively). Codecs were also evaluated at their maximum reconstruction power B_{Max}, obtained similarly by performing an as low as possible compression in the workflow. The evaluation was carried out in three separate steps, to better assess the impact of the compression and the rendering in the final result:

1. B against B_{Ref}: Evaluation of the combined impact of encoder, decoder and renderer of the proposed algorithm against the uncompressed rendered content, on four fixed compression ratios.

2. B against B_{Max}: Evaluation of the impact of encoder and decoder of the proposed algorithm, using as reference the results of running the encoder at its maximum reconstruction quality B_{Max}. This step was implemented to isolate the impact of the proposed renderer on the overall quality.

3. B_{Max} against B_{Ref}: Evaluation of the proposed renderer with respect to the reference renderer. This step was implemented to assess the proposed rendering model without the influence of compression artefacts.

All three evaluation steps were implemented for the objective assessment, whereas for the subjective assessment the second step was discarded, as changing the reference from B_{Ref} to B_{Max} in the tests would have biased the results.

Objective quality metrics

To evaluate the impact of the distorsions caused by the proposed algorithms, PSNR and SSIM were selected from the literature to objectively assess the visual quality of the contents. The metrics were applied separately to the luma channel Y and for each viewpoint image, as follows:

$$PSNR_Y(k,l) = 10\log_{10}\frac{1023^2}{MSE(k,l)}, \tag{6.8}$$

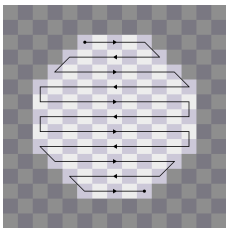


Figure 6.13 – Ordering of the views for the subjective tests for the lenslet case. ©2018 IEEE

$$SSIM_Y(k,l) = \frac{(2\mu_I\mu_R + c_1)(2\sigma_{IR} + c_2)}{(\mu_I^2 + \mu_R^2 + c_1)(\sigma_I^2 + \sigma_R^2 + c_2)}, \tag{6.9}$$

in which k and l are the indexes of the perspective views, $MSE(k,l)$ is the mean square error, μ_I and μ_R are the mean values, σ_I^2 and σ_R^2 are the variances, and σ_{IR} is the covariance of the two perspective views in channel Y. Please note that, unlike in the previous section, we are now computing the PSNR and SSIM on 10-bit data. PSNR was computed for chrominance channels U, V following Equation 6.8, and a weighted average [Ohm et al., 2012] was calculated as follows:

$$PSNR_{YUV}(k,l) = \frac{6PSNR_Y(k,l) + PSNR_U(k,l) + PSNR_V(k,l)}{8} \tag{6.10}$$

The average PSNR value for Y channel was then computed across the viewpoint images:

$$\widehat{PSNR}_Y = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} PSNR_Y(k,l), \tag{6.11}$$

in which K and L represent the number of perspective views. For the lenslet case, $K = L = 13$, as the outermost perspective viewswere deemed too distorted to be used in the evaluation, whereas for the HDCA case, $K = 101$ and $L = 21$. $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ were analogously computed following Equation 6.11.

Subjective Methodology

Following the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012], a comparison-based adjectival categorical judgement methodology with a 7-point grading scale was selected to perform the subjective visual quality assessment, from -3 (much worse) to +3 (much better),

with 0 indicating no preference.

A passive assessment was considered in order to ensure the same experience for all participants, following what has been said in Chapter 3. For the lenslet case, participants were shown the light field contents as pre-recorded animations navigating between the perspective views in a serpentine order, to mimic the parallax effect (Figure 6.13). To avoid negative bias in the subjects, only a subset of 97 out of 225 perspective views was presented in the animation, as we did in Chapter 3.2, since the rest of the views already presents high visual distorsion before compression that can negatively affect the results. The views were displayed at a rate of 10 frames per second (fps), to ensure a smooth transition. The total length of the animation was 9.7 seconds. Each stimulus was displayed alongside the uncompressed reference in a side-by-side arrangement. The position of the reference was fixed for the duration of the test, and participants were informed beforehand on which side of the screen the reference would be displayed.

For the HDCA case, a subset of views was selected for the visualization, as rendering the entire set would have led to an excessive length per stimulus. Both horizontal, vertical and diagonal movements were included in the passive representation to aid in recognizing compression artifacts through motion disparity. In order to properly show the contents side by side with the reference, each perspective view was cropped to half the size of the screen. The views were displayed at a rate of 30 frames per second (fps), as a lower framerate would have led to jerkiness in visualization. The total length of the animation was 12 seconds. Each stimulus was displayed alongside the uncompressed reference in a side-by-side arrangement. The position of the reference was fixed for the duration of the test, and participants were informed beforehand on which side of the screen the reference would be displayed.

Participants were asked to rate the quality of the test stimuli when compared to the uncompressed reference. A training session was organized before the experiment to familiarize participants with artefacts and distorsions in the test images. Four training samples, created by compressing one additional content from the dataset on various bitrates, were manually selected by expert viewers. In the lenslet case, the experiment was split in four sessions. In each session, 46 stimuli (47 in the last session) were shown along with the uncompressed reference, corresponding to approximately 8 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all sessions, hence evaluating all 185 stimuli. A break of ten minutes was enforced between sessions. In the HDCA case, only one session displaying the 60 stimuli was employed.

The test was conducted in a laboratory for subjective video quality assessment, which was set up according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. A professional Eizo ColorEdge CG318-4K 31.1-inch monitor with 10-bit depth and native resolution of 4096×2160 pixels was used for the tests. The monitor settings were adjusted according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black

level of 0.2 cd/m^2. The controlled lighting system in the room consisted of adjustable neon lamps with 6500 K color temperature, while the color of the background walls was mid grey. The illumination level measured on the screens was 15 lux. The distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content, conforming to requirements in ITU-R Recommendation BT.2022 [ITU-R BT.2022, 2012]. Subjects were allowed to move further or get closer to the screen.

For the lenslet case, a total of 28 subjects (19 males and 9 females) participated in the test, whereas for the HDCA case, 20 subjects (6 males and 14 females) took part in the test, for a total of 28 and 20 scores per stimulus, respectively. Subjects were between 18 and 35, with a mean age of 23.14 years old for the lenslet case and 23.11 for the HDCA case. Before starting the test, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively.

Subjective Data Processing and Statistical Analysis

Outlier detection and removal was conducted on the collected scores, according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. No outlier was detected, leading to 28 scores per stimulus. The MOS was computed for each stimulus, and the corresponding 95% CIs were calculated assuming a Student's t-distribution.

To determine whether the differences in MOS between the proponents were statistically different, a one-sided Welch's test at 5% significance level was conducted on the results, with the following hypotheses:

$$H0 : MOS_A \leq MOS_B$$
$$H1 : MOS_A > MOS_B,$$

in which A and B are the proposed algorithms under comparison. The test was conducted for each compression ratio and for each content. If the null hypothesis were to be rejected, then it could be concluded that codec A performed better than codec B for the given content and compression ratio, at a 5% significance level. Additionally, a one-way ANOVA test was performed on the results to determine the overall difference between codecs.

6.2.3 Results

Lenslet camera

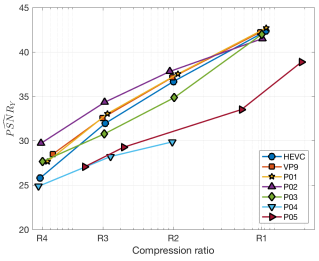
Figures 6.14, 6.15 and 6.16 show the results of the objective evaluation campaign for the lenslet case. Results of $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ computed using B_{Ref} as reference (Figure 6.14) show that all codecs have similar performance for compression ratio $R1$, with the exception of $P05$ and $P06$, which were considerably worse. For compression ratios $R2$ and $R3$, codecs $P04$ and $P05$ perform worse than the other codecs, while $P01$ and $P02$ achieve the best results. In particular, $P01$ and VP9 have similar performance, whereas HEVC has a slightly poorer behaviour. For the lowest bitrate, $P02$ clearly outperforms the anchors and other codecs.

The comparison of the results obtained using B_{Max} as reference (Figure 6.15) exhibit similar trends, although $P02$ shows a significant gain in performance when using $\widehat{PSNR}_Y$ as metric. It is worth mentioning that, in the $\widehat{PSNR}_Y$ case, proposal $P03$ seems to perform consistently worse when the reference is set to B_{Max} when compared to reference B_{Ref} (Figure 6.15), at least for higher bitrates. This is likely due to color space conversions negatively affecting the performance of JPEG 2000, used in the codec. The comparison of B_{Max} against B_{Ref} (Figure 6.16) shows that all proposed renderers achieve favorable results, with the exception of $P04$. This is mainly due to the fact that the codec uses a mixture of Gaussians to represent the light field structure, leading to poor results when using full-reference objective quality metrics.

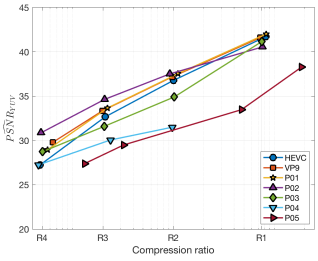
Figures 6.17 and 6.18 present the outcome of the subjective evaluation campaign for the lenslet case. Results show similar performance for all codecs in highest bitrate, with the exception of $P05$ and $P06$ (Figure 6.17 (a - e) and Figure 6.18 (d)). Among all proponents, $P01$ has the best performance, $P02$ being a close second. For compression ratio $R2$, proponents $P01$ and $P02$ perform similar to anchor VP9 and they surpass the other codecs on more than three out of five contents (Figure 6.18 (c)). The same trend can be observed for compression ratio $R3$, where $P01$ is never outplayed and always performs better than the other codecs, with the exception of $P02$, which has worse results for only one out of five contents (Figure 6.18 (b)). For the lowest bitrate, $P02$ has the best performance, ranking better than the other codecs on at least three out of five contents, followed by $P03$ and $P01$ (Figure 6.18 (a)).

Subjective results show that B_{Max} is never perceived as better than B_{Ref}, and in certain cases it is considered as significantly worse than the reference (Figure 6.17 (f)). In particular, while some proposed renderers were sometimes rated as slighly better than the reference, they fail to be significantly better, as the CIs are always seen to be crossing the zero. Moreover, in case of content $I05$, only $P01$ and $P02$ are considered equivalent to the reference, while all other codecs significantly underperform when compared to the reference renderer. Additionally, the renderer proposed in $P04$ is always perceived as worse than the reference, mainly due to the blur caused by the Gaussian model.

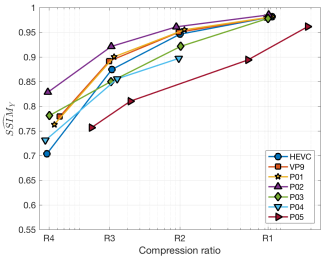
One-way ANOVA performed on the results of the subjective tests confirms that the codecs



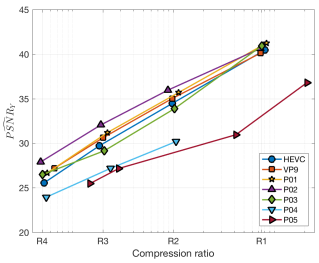
(a) I01, $\widehat{PSNR}_Y$, B against B_{Ref}



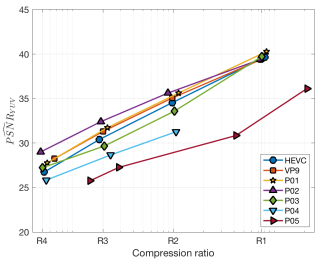
(b) I01, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



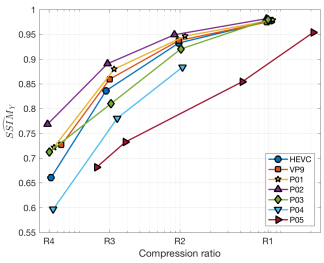
(c) I01, $\widehat{SSIM}_Y$, B against B_{Ref}



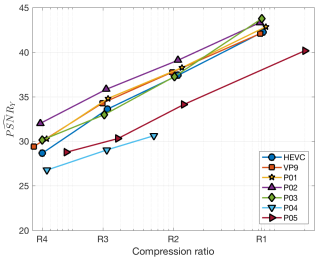
(d) I02, $\widehat{PSNR}_Y$, B against B_{Ref}



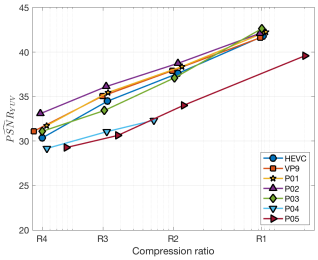
(e) I02, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



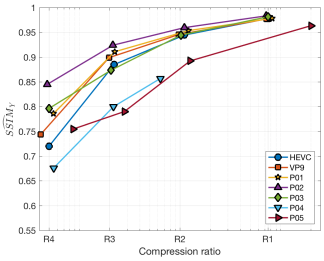
(f) I02, $\widehat{SSIM}_Y$, B against B_{Ref}



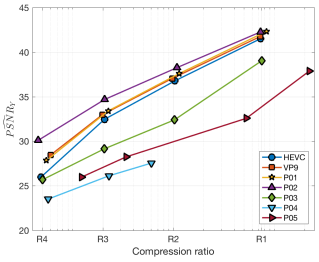
(g) I04, $\widehat{PSNR}_Y$, B against B_{Ref}



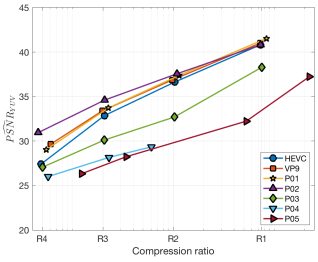
(h) I04, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



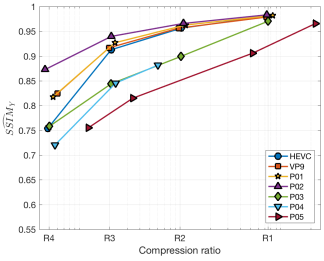
(i) I04, $\widehat{SSIM}_Y$, B against B_{Ref}



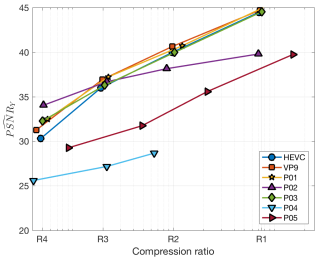
(j) I09, $\widehat{PSNR}_Y$, B against B_{Ref}



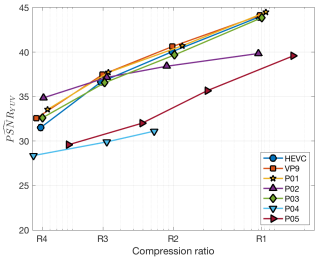
(k) I09, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



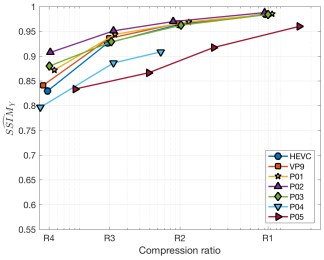
(l) I09, $\widehat{SSIM}_Y$, B against B_{Ref}



(m) I10, $\widehat{PSNR}_Y$, B against B_{Ref}

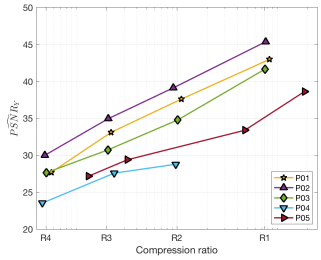


(n) I10, $\widehat{PSNR}_{YUV}$, B against B_{Ref}

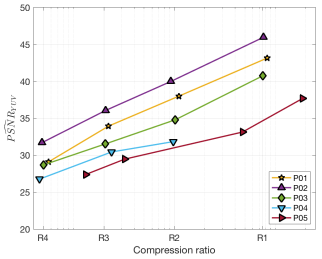


(o) I10, $\widehat{SSIM}_Y$, B against B_{Ref}

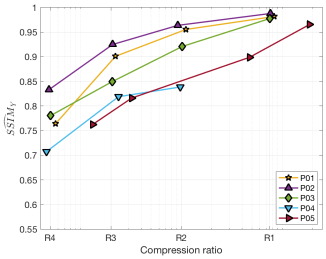
Figure 6.14 – Results of the objective evaluations comparing B against B_{Ref}, for all lenslet contents (rows). $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.



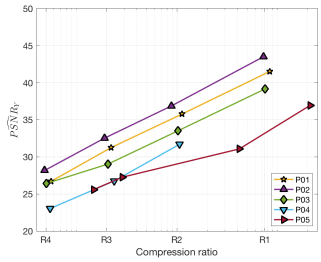
(a) I01, $\widehat{PSNR}_Y$, B against B_{Max}



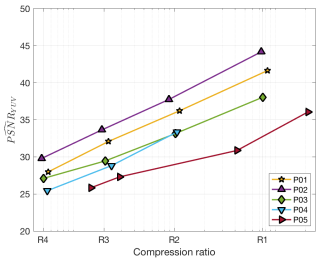
(b) I01, $\widehat{PSNR}_{YUV}$, B against B_{Max}



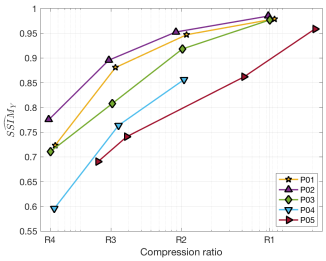
(c) I01, $\widehat{SSIM}_Y$, B against B_{Max}



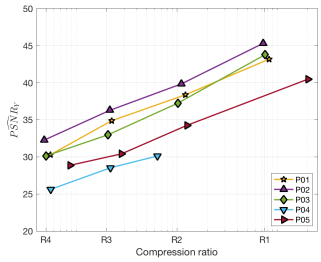
(d) I02, $\widehat{PSNR}_Y$, B against B_{Max}



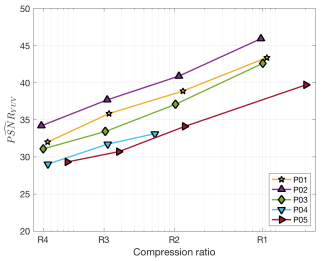
(e) I02, $\widehat{PSNR}_{YUV}$, B against B_{Max}



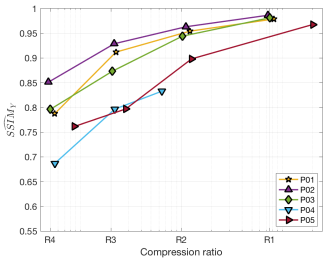
(f) I02, $\widehat{SSIM}_Y$, B against B_{Max}



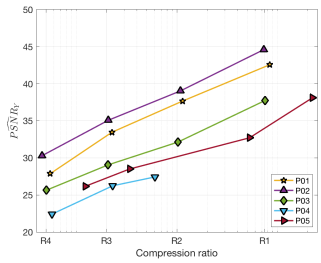
(g) I04, $\widehat{PSNR}_Y$, B against B_{Max}



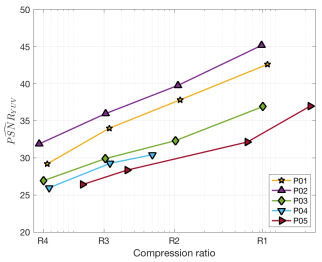
(h) I04, $\widehat{PSNR}_{YUV}$, B against B_{Max}



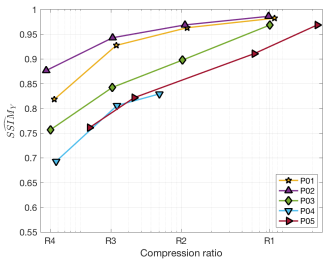
(i) I04, $\widehat{SSIM}_Y$, B against B_{Max}



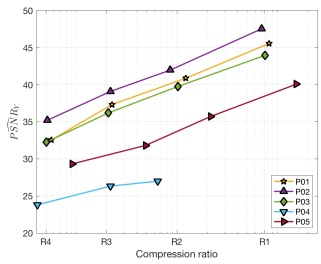
(j) I09, $\widehat{PSNR}_Y$, B against B_{Max}



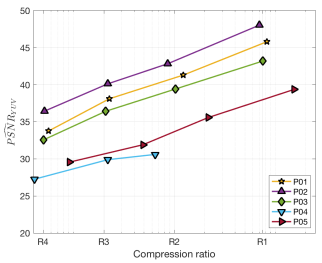
(k) I09, $\widehat{PSNR}_{YUV}$, B against B_{Max}



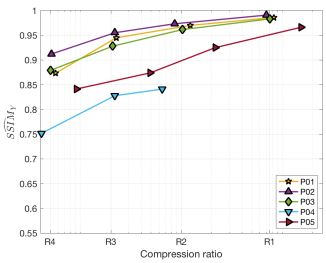
(l) I09, $\widehat{SSIM}_Y$, B against B_{Max}



(m) I10, $\widehat{PSNR}_Y$, B against B_{Max}

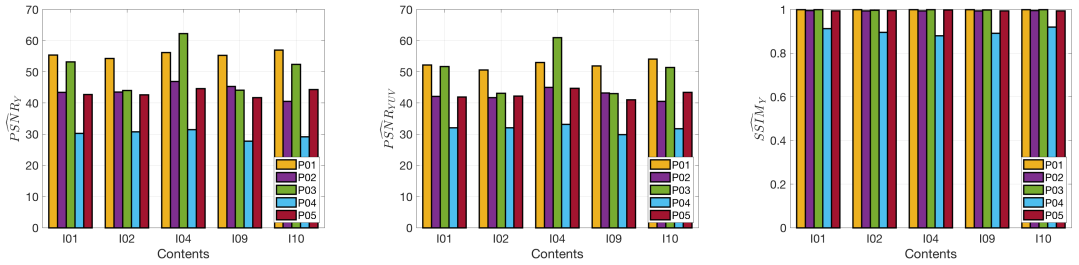


(n) I10, $\widehat{PSNR}_{YUV}$, B against B_{Max}



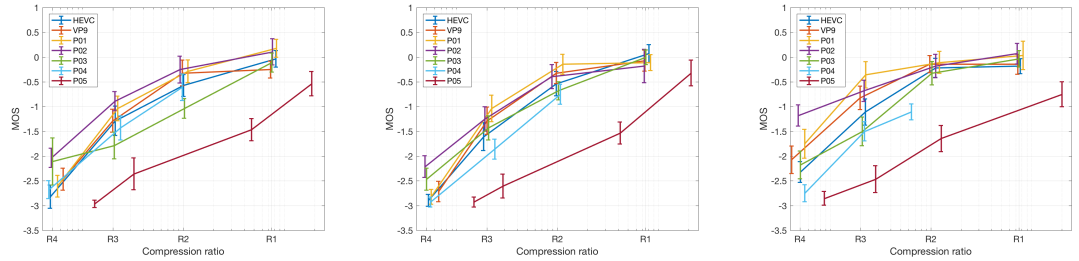
(o) I10, $\widehat{SSIM}_Y$, B against B_{Max}

Figure 6.15 – Results of the objective evaluations comparing B against B_{Max}, for all lenslet contents (rows). $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.

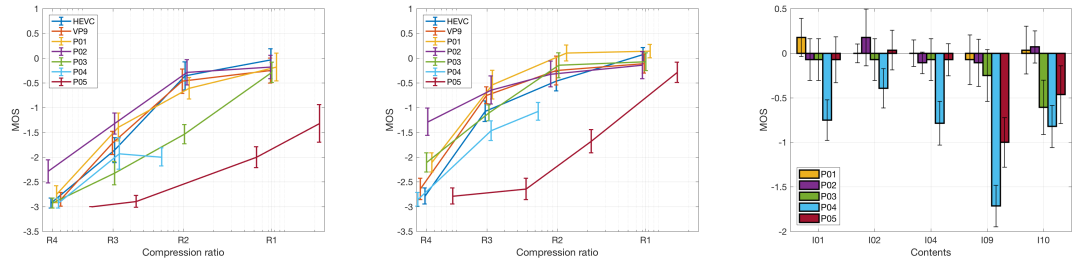


(a) $\widehat{PSNR}_Y$, B_{Max} against B_{Ref} (b) $\widehat{PSNR}_{YUV}$, B_{Max} against B_{Ref} (c) $\widehat{SSIM}_Y$, B_{Max} against B_{Ref}

Figure 6.16 – Results of the objective evaluation, comparing B_{Max} with respect to B_{Ref} for all lenslet contents. $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.

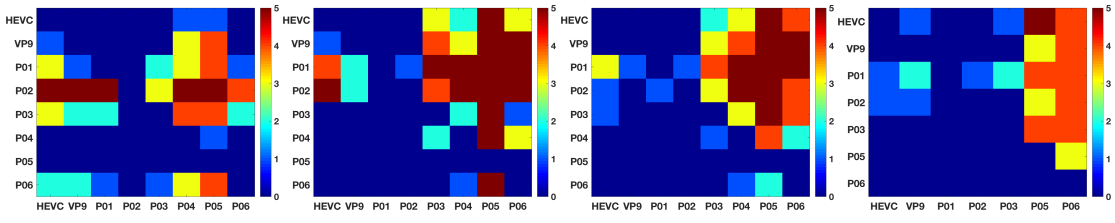


(a) I01, B against B_{Ref} (b) I02, B against B_{Ref} (c) I04, B against B_{Ref}



(d) I09, B against B_{Ref} (e) I10, B against B_{Ref} (f) B_{Max} against B_{Ref}

Figure 6.17 – Results of the subjective evaluation. MOS vs bitrate, with respective CIs (a - e), and comparison of B_{Max} with respect to B_{Ref} (f), for all lenslet contents.



(a) R4 (b) R3 (c) R2 (d) R1

Figure 6.18 – Pairwise comparison results for subjective tests in the lenslet case. Each cell contains the number of contents for which the null hypothesis was rejected, for each compression ratio. The null hypothesis is defined as $MOS_i \leq MOS_j$, in which i indicates the row and j the column of the matrix.

are significantly different ($p = 2.1376 \times 10^{-111}$). In particular, proponent $P03$ has comparable performance with respect to the anchors. Proponents $P01$ and $P02$ have statistically equivalent behaviour, whereas they are statistically better than the anchors. On the other hand, $P04$, $P05$ and $P06$ perform statistically worse than the anchors.

High density camera array

Figure 6.19 shows the results of the objective quality evaluation for HDCA contents. Results from the comparison between B and B_{Max} are omitted as exhibiting very similar behavior with respect to the comparison between B and B_{Ref}. It can immediately be observed that for low bitrates proponent $P06$ is performing consistently better than the anchors, according to all the metrics. However, when it comes to high bitrates, the results are strongly dependent on the content. In particular, for content $S06$ gains can be observed for all the metrics, but most notably for $\widehat{SSIM}_Y$; on the other hand, the anchors perform remarkably better than $P06$ on at least 2 out ot four contents when using $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$ (see Figure 6.19 (g-h) and (j-k)).

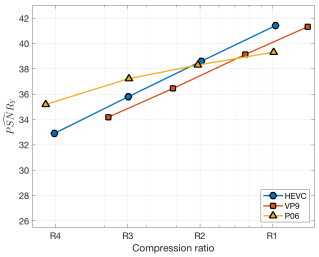
Figures 6.20 and 6.21 show the results of subjective quality evaluation. Results confirm that, for low bitrates, proponent $P06$ performs significantly better than the anchors (see Figure 6.21(a-b), in which $P06$ outperforms the anchors on at least 3 out of 4 contents). However, for high bitrates, the proposed solution is either equivalent or worse than the anchors, failing to be significantly better (see Figure 6.21(c-d)). Moreover, for the highest bitrate $P06$ is outperformed by HEVC on half the contents.

Results from the comparison between the reference renderer B_{Ref} and the proposed MR-DIBR show that, in 3 out of 4 cases, the two are statistically equivalent (see Figure 6.20(a-c)). It is worth noting that for one content it appears to be statistically better than the reference renderer. The results show that MR-DIBR is a valuable alternative to synthetize the perspective views at the receiver side, since no loss in quality can be perceived in absence of compression artifacts.

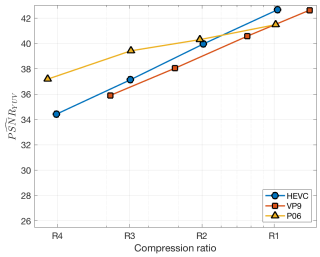
6.3 Conclusions

In this chapter we have reported the results of objective and subjective quality assessment performed under the framework of two grand challenges for compression of light field contents. Our contributions are the following:

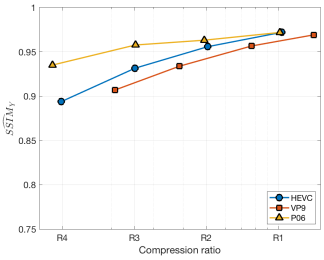
- We first conducted the evaluation campaign for the ICME 2016 Grand Challenge on Light Field Compression. We applied state-of-the-art objective quality metrics and designed the subjective evaluation methodology to properly assess the performance of the five proposed solutions for compression of lenslet-based light field images. We



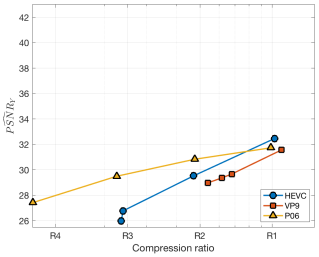
(a) S02, $\widehat{PSNR}_Y$, B against B_{Ref}



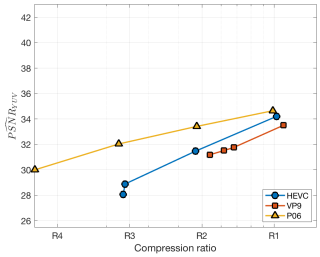
(b) S02, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



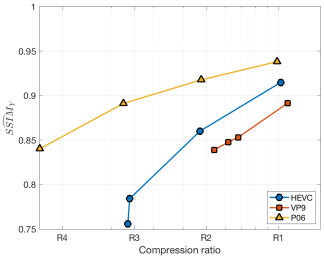
(c) S02, $\widehat{SSIM}_Y$, B against B_{Ref}



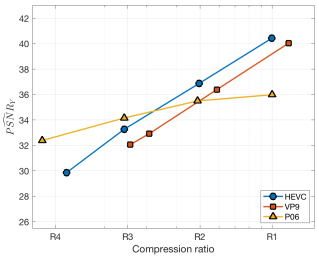
(d) S06, $\widehat{PSNR}_Y$, B against B_{Ref}



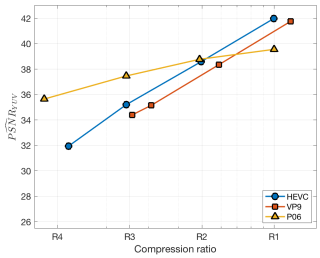
(e) S06, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



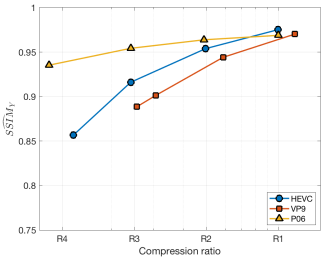
(f) S06, $\widehat{SSIM}_Y$, B against B_{Ref}



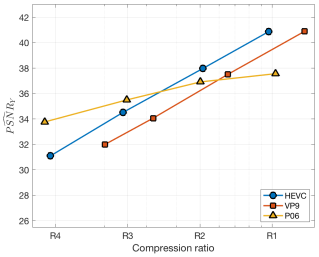
(g) S09, $\widehat{PSNR}_Y$, B against B_{Ref}



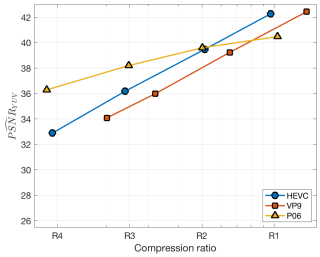
(h) S09, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



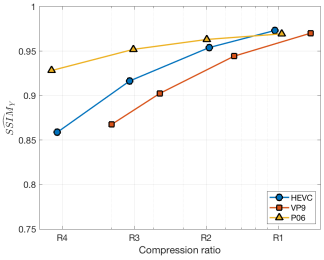
(i) S09, $\widehat{SSIM}_Y$, B against B_{Ref}



(j) S10, $\widehat{PSNR}_Y$, B against B_{Ref}

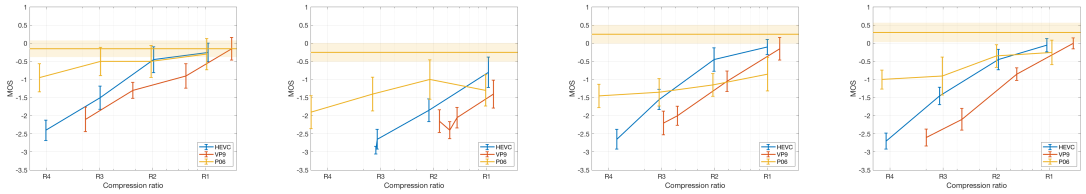


(k) S10, $\widehat{PSNR}_{YUV}$, B against B_{Ref}



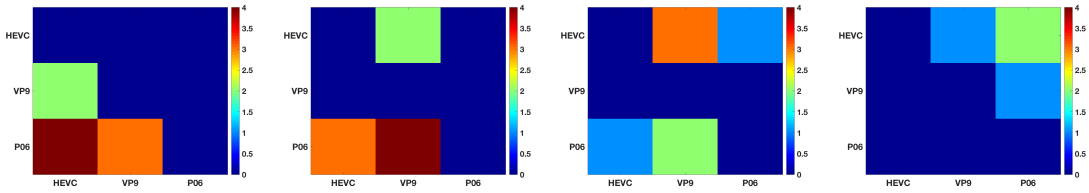
(l) S10, $\widehat{SSIM}_Y$, B against B_{Ref}

Figure 6.19 – Results of the objective evaluations comparing B against B_{Ref}, for all HDCA contents (rows). $\widehat{PSNR}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ are used as metric in the first, second and third column, respectively.



(a) S02 (b) S06 (c) S09 (d) S10

Figure 6.20 – Results of the subjective evaluation. MOS vs bitrate, with respective CIs, and comparison of B_{Max} with respect to B_{Ref} (shaded) for all HDCA contents.



(a) R4 (b) R3 (c) R2 (d) R1

Figure 6.21 – Pairwise comparison results for subjective tests in the HDCA case. Each cell contains the number of contents for which the null hypothesis was rejected, for each compression ratio. The null hypothesis is defined as $MOS_i \leq MOS_j$, in which i indicates the row and j the column of the matrix.

showed that there is much to be gained in using new compression schemes as opposed to legacy JPEG.

- We then designed the evaluation procedure for the ICIP 2017 Grand Challenge on Light Field Coding, held concurrently with the JPEG Call for Proposals on Light Field Coding. In particular, we performed the objective and subjective quality assessment on various steps, to analyze the performance of both encoders and renderers that were proposed as a response. Two light field acquisition techniques were considered as input. Results showed that direct application of state-of-the-art video codecs to compress light field images can be improved using new codec designs. In particular, two codecs were found to outperform others in both objective and subjective terms. We also demonstrated that no proposed representation model is statistically better than that adopted as reference.

It should be remarked that the assessment conducted in this chapter only takes into consideration perceptual quality when proclaiming a winner. In addition to compression efficiency and visual quality, other criteria such as complexity, delay and random access should be also considered when adopting a preferred solution.

7 Impact of coding approaches on compression efficiency for light field images

Disclaimer: Some of the contents of this chapter were adapted from the following article, with permission from all co-authors and publishing entities:

Viola, Irene, Martin Řeřábek, and Touradj Ebrahimi. "Comparison and evaluation of light field image coding approaches" in IEEE Journal of selected topics in signal processing 11.7 (2017): 1092-1106. ©2017 IEEE.

Personal contribution: I was the main curator of the experiments, from the selction of the codecs to the design of the quality evaluation. I performed the tests and conducted the analysis of the results.

As large volumes of data are generated in the acquisition of light field contents, efficient compression solutions are needed to reduce the burden on both transmission and storage systems. Still, finding a suitable compression algorithm for light field contents is largely an open problem. Several publications have been devoted in the literature to propose new coding techniques aimed at exploiting the intrinsic redundancy of light field information. However, based on the acquisition technology employed to capture the light field data, different solutions may be needed to effectively reduce the amount of data while preserving an acceptable quality level.

Currently available techniques to capture and visualize light field images determine two general approaches for light field image compression. A general diagram of workflow for light field image acquisition and visualization is depicted in Figure 7.1. The first coding approach assumes that the raw sensor data obtained during the acquisition step is compressed directly with basic signal pre-processing such as demosaicing or devignetting (point A in Figure 7.1). The actual format of raw data strongly depends on the exact acquisition device, e.g. a lenslet based hand-held camera, a multi-camera array, or a multi-view plus depth acquisition device. Often, extensive post-processing of the decompressed light field image is necessary prior to its visualization. Furthermore, additional metadata about the captured scene and acquisition device, e.g. camera and color calibration data, is needed to properly process and visualize the

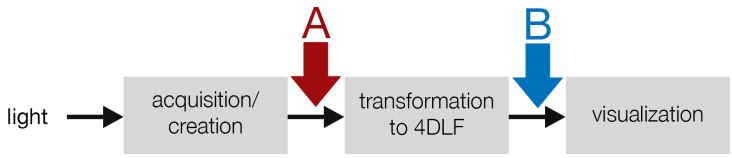


Figure 7.1 – General acquisition and display pipeline for light field images. ©2017 IEEE

light field image.

The second coding approach considers creating a 4D data structure from the light field image prior to compression (point B in Figure 7.1). The 4D light field is composed by a collection of perspective images, which can be visualized without a need for acquisition related metadata or post-processing. The process of creating the 4D light field from the raw sensor data strongly depends on the exact acquisition device.

In the context of general light field image manipulation, one can think of two specific use cases related to either of the two coding approaches defined above. On the one hand, professional photographers, operators, and artists may benefit from light field image acquisition technologies, since they allow for greater flexibility in terms of optimal parameters selection after capture. For example, an erroneous selection of focal plane in a scene may lead to several retakes and thus to greater costs. Other features, such as change of point of view or zoom, may dramatically impact the way scenes are captured. In this case, it is of paramount importance that key factors in the acquisition, such as white balance, color, and exposition, are not altered in the compression step, and that acquisition metadata is stored to be used during post-processing.

Consumers, on the other hand, may turn to light field imaging when looking for an enhanced experience to capture a special moment. Ability to change zoom, perspective, and focus in a simple and intuitive way without the need for expensive post-processing software, is in line with the interactivity already seen in applications like Instagram, in which users can modify the appearance of the captured scenes with predefined filters. In this case, the fidelity to the acquisition parameters is less important. However, the resulting image should not be too large and ready to be visualized and shared in devices with limited resources.

In this chapter, we compare two coding approaches to compress light field images. The first performs the compression on the minimally pre-processed raw data (point A in Figure 7.1). As such, it is strongly acquisition dependent, since the compression solution will be tailored on the way the data is captured and arranged. Moreover, perceptualy-based solutions will be harder to employ, since the data needs to be heavily processed (often with the aid of metadata information) to be rendered. The second approach enforces the transformation to a 4D data structure of perspective views (4D light field) prior to the compression. As the coding is to be considered mostly independent from the acquisition step, any additional information carried by the raw data should be either discarded or separately accounted for. To aid in our evaluation, five different compression algorithms, which employ either of the two approaches

Figure 7.2 – Processing chain for lenslet image compression used for two compression algorithms (anchor *P*01, proponent *P*02). ©2017 IEEE

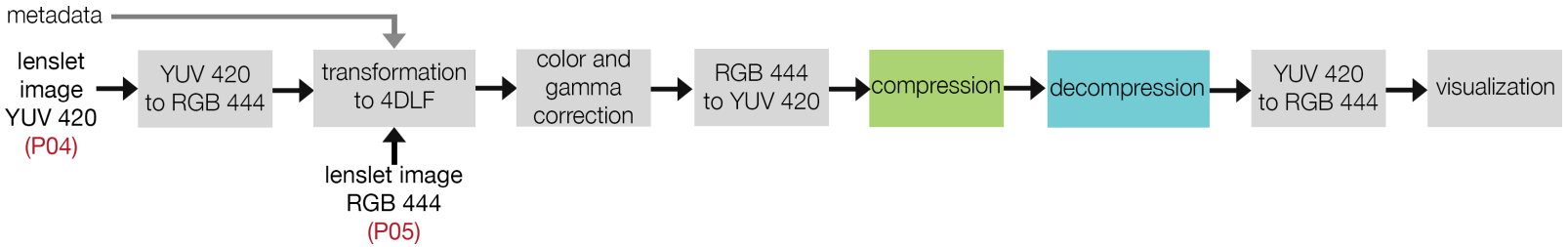


Figure 7.3 – Processing chain for 4D light field compression used for two compression algorithms (anchors *P*04 and *P*05). ©2017 IEEE

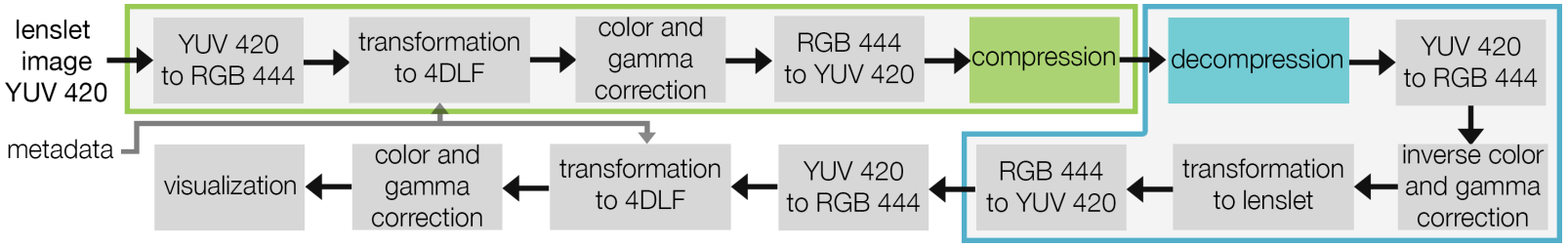


Figure 7.4 – Processing chain for hybrid compression of lenslet using intermediate 4D light field transformation (proponent *P*03). The green and blue blocks highlight how the compression step involves intermediate transformation to 4D light field, and the decompression step involves the inverse transformation to lenslet image. ©2017 IEEE

and are suitable for the predefined use cases mentioned before, are described, investigated, and evaluated through a set of objective and subjective quality evaluations. For the objective quality assessment, the PSNR metric is used, while for the subjective evaluation, the two methodologies described in Chapter 3.2 have been selected. The first methodology allows for interaction with the displayed content, whereas the second passively shows the different perspective views composing the light field content, in order to ensure that all users see and rate the exact same content. We present and compare the results of the different coding approaches, drawing conclusions and recommendations for the design of future compression solutions.

7.1 Light field coding strategies

This section describes in details the coding approaches investigated in the evaluation process, including a thorough description of five selected algorithms to compress light field images.

Two main coding approaches can be considered for compression of light field images. Referring to the general diagram of workflow presented in Figure 7.1, we can compress the data at two different stages. Compression can be performed on the raw sensor data that has been captured with the selected acquisition technology, after minimal processing, such as demosaicing and devignetting (point A in Figure 7.1). The 4D light field can be recovered from the decompressed bitstream through extensive post-processing, involving camera and color calibration metadata which needs to be sent along with the bitstream. The second coding approach performs compression on the 4D light field obtained from the raw data (point B in Figure 7.1). The 4D light field is a collection of perspective views which can be visualized as they are, or combined to create new interpolated views, synthetic aperture, refocusing, and extended focus. Since the transformation of raw sensor image data to 4D light field is performed before the compression step, no metadata is required for visualization. The compression solutions used to code the raw sensor data, as well as the transformation to 4D light field from the raw sensor data, strongly depend on the acquisition technology used to capture light field images. If compression is applied at point A, the selected scheme will profoundly differ based on the acquisition technology. On the other hand, a compression scheme operating at point B can compress 4D light field image information captured with any acquisition technology.

In order to compare the two coding approaches on a common ground, we decided to focus our attention on evaluating coding strategies for lenslet-based acquisition. Lenslet-based acquisition allows to compare the two approaches on the same image content captured within the same conditions. In this case, the raw sensor data is minimally pre-processed to obtain a lenslet image. From the lenslet image, the 4D light field can be recovered through rectification, calibration and extraction of perspective images, using camera and color calibration data. The extraction of perspective images from the lenslet image generates $N \times M$ views, depending on the uv resolution. However, the most external views contain too many distortions to be

Table 7.1 – Summary of compression schemes. ©2017 IEEE

Proponent	Description
P01	Lenslet image compressed using HEVC intra (software x265).
P02	Lenslet image compressed using HEVC intra with LLE and SS (software HM-14.0) [Monteiro et al., 2016].
P03	Lenslet image compressed using intermediate transformation to perspective views and HEVC (software JEM 2.0) [Liu et al., 2016].
P04	Chroma subsampling of the lenslet image and compression of perspective views through pseudo-temporal sequence using HEVC (software x265).
P05	Compression of perspective views through pseudo-temporal sequence using HEVC (software x265).

properly visualized. The 4D light field coding approach can take advantage of this fact by not coding the most distorted views, which will likely not be used in the visualization process, thus further reducing the size of the bitstream.

The rest of the section is organized as follows. The first coding approach, which deals with compression of lenslet images, is described. Then, the second coding approach, which focuses on compression of 4D light field obtained from lenslet images, is presented. Finally, one hybrid approach to compress lenslet images through transformation to 4D light field, introduced in ICME 2016 Grand Challenge, is detailed. Authors are aware of practical drawbacks and flaws in this solution. However, it was decided to include it in the evaluation process because of its optimal performance within the Grand Challenge, and because it represents a transition point between the two coding approaches. A summary of the compression schemes can be found in Table 7.1.

7.1.1 Lenslet image compression

The lenslet coding approach performs compression on the lenslet image, obtained from the raw sensor data after demosaicing and devignetting. Figure 7.2 depicts the workflow for the coding approach. The workflow was adopted following the definition of the ICME 2016 Grand Challenge, which required to perform compression on YUV 420 lenslet images in 8-bit precision.

The raw sensor data is first demosaiced, devignetted and clipped to 8 bits to obtain a lenslet image. The lenslet image is subsequently converted to YUV 420 format, and compressed and decompressed using the selected compression scheme. The output of the decompression step is then upsampled and converted to RGB 444. Conversion to RGB 444 format is required to perform the transformation from lenslet image to 4D light field. The 4D light field is created from the decompressed lenslet image using camera metadata. Color and gamma corrections are applied separately on each view. The perspective views forming the 4D light field can

subsequentially be visualized on commercially available displays, or combined to create new interpolated views, synthetic aperture or refocusing effect.

Two compression algorithms (*P*01 and *P*02) compliant with the workflow depicted in Figure 7.2 were evaluated. One compression scheme was selected among the best performing submitted to ICME 2016 Grand Challenge, and it was compared to an HEVC anchor. The anchor *P*01 exploits HEVC intra profile with default settings to compress the YUV 420 lenslet image. To perform the compression, the software x265 was used[1]. The second algorithm *P*02 uses a modified version of HEVC intra profile, which integrates LLE and SS to exploit the redundancies in the lenslet structure [Monteiro et al., 2016]. The image is partitioned into blocks using HEVC intra prediction scheme. LLE estimates the current block by selecting the best linear combination of k nearest neighbors through a least-square optimization problem. SS predicts the current block using the best among two blocks, one given by best block matching in the search window, the other chosen by searching for best linear combination between the first selected block and another block in the search window. The codec performs both LLE and SS, and then choses the prediction method that gives the smallest rate distortion cost.

7.1.2 4D light field compression

The 4D light field coding approach performs the compression on the 4D light field, obtained from the lenslet image, after color and gamma corrections. Figure 7.3 depicts the workflow for this coding approach. Two anchors were created to assess the visual quality of this coding approach. The first anchor *P*04 assumes the same input as the compression schemes using the lenslet coding approach (YUV 420 lenslet images). The color space is then upsampled and converted again to RGB 444, to be used in the transformation process. To assess the effect of chroma subsampling of the lenslet image on the resulting quality of the final 4D light field, the second anchor *P*05 performs the compression on the 4D light field, obtained from lenslet in RGB 444 format, after color and gamma corrections (Figure 7.3). In this case, the lenslet image is not transformed to YUV from RGB, and the color space is not subsampled before the transformation.

For both anchors, the 4D light field is created from the uncompressed lenslet image using camera metadata, and color and gamma corrections are applied separately on each view, prior to compression. Each view is converted from RGB 444 to YUV 420. The views are arranged in a pseudo-temporal sequence in spiral order, as depicted in Figure 7.6. Due to the geometrical distortions present in the most external views, only a subset of the views is coded. Specifically, only the 13×13 internal views out of 15×15 views are encoded. The pseudo-temporal sequence is coded with HEVC software x265. In the decompression step, the views which have not been coded are replaced with copies of neighboring views, to reconstruct the 15×15 images that compose the 4D light field. After decompression, the views are upsampled and converted

[1] https://www.videolan.org/developers/x265.html

(a) Bikes (b) Stone_Pillars_Outside (c) Fountain_&_Vincent_2 (d) Friends_1

Figure 7.5 – Central perspective view from each content used in our experiment. ©2017 IEEE

to RGB 444 and rearranged in the 4D light field. The perspective images composing the 4D light field can then be visualized on commercially available displays, or combined to create synthetic aperture, refocusing and new interpolated views.

7.1.3 Hybrid compression of lenslet images

Among the participants to the ICME 2016 Grand Challenge, which required to have YUV 420 lenslet image as input and output of the compression and decompression step, one algorithm performed compression on lenslet images using intermediate transformation to 4D light field [Liu et al., 2016]. Figure 7.4 depicts the workflow for this algorithm.

The algorithm *P*03 proposes a compression of 4D light field images based on pseudo-sequences of perspective views. Due to the constraints of the Grand Challenge, the YUV 420 lenslet image is first converted to RGB 444 color space, to be used in the transformation step. Then the lenslet is processed to obtain the perspective views that compose the 4D light field. The views are color and gamma corrected and then converted back to YUV 420. A subset of them is then rearranged in a specific coding order, that accounts for similarities between adjacent views, and coded using the JEM encoder[2]. In the decompression step, the views are rearranged in the 4D light field. Inverse color and gamma corrections are applied and the lenslet image is formed following the inverse process of the transformation to 4D light field.

The conversion from lenslet to 4D light field and back was needed to be compliant with the requirements of the grand challenge. However, it can be clearly seen that the proposed approach is hybrid, in the sense that it compresses lenslet images through transformation to 4D light field. The tranformation from lenslet images to 4D light field and back is lossless, as it is defined in [Liu et al., 2016].

7.2 Experiment design

This section describes the evaluation process in details. First, the data preparation process is presented, along with the coding conditions. Methodologies and metrics for objective and

[2]https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/tags/HM-16.6-JEM-2.0rc1/

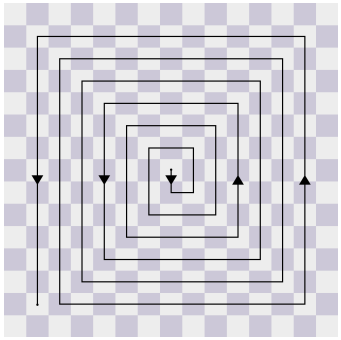


Figure 7.6 – Ordering of the views for coding. ©2017 IEEE

subjective evaluation are then presented in details.

7.2.1 Dataset preparation and coding conditions

Four light field images, acquired by a Lytro Illum camera, were selected from the publicly available EPFL light field image dataset [Řeřábek and Ebrahimi, 2016]. More specifically, *Bikes*, *Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* contents were selected for our experiments. The central view of each content used is depicted in Figure 7.5. The images were carefully selected from those used in the ICME 2016 Grand Challenge [Viola et al., 2016a] in order to provide a wide range of scenarios, containing details that would prove challenging for the compression algorithms. To obtain the 4D light field, the lenslet images were processed using the light field MATLAB toolbox [Dansereau et al., 2013] [Dansereau et al., 2015].

The compression algorithms were evaluated on four bitrates (corresponding to four compression ratios), namely $R1 = 1$ bpp $(10:1)$, $R2 = 0.5$ bpp $(20:1)$, $R3 = 0.25$ bpp $(40:1)$, $R4 = 0.1$ bpp $(100:1)$. The compression ratios are computed as ratios between the size of the uncompressed raw images in 10bit precision ($5368 \times 7728 \times 10$ bits $= 414839040$ bits) and the size of the compressed bitstreams.

The uncompressed reference was obtained by demosaicing, devignetting and clipping to 8 bits the raw sensor data, transforming it to 4D light field and applying color and gamma corrections. Unlike the reference used in ICME 2016 Grand Challenge, which used as a reference the 4D light field obtained from YUV 420 lenslet image, we obtain our reference from the lenslet image in RGB 444, without any chroma subsampling. This reference was selected to have a proper comparison with acquisition data obtained with minimal pre-processing. For this reason, chroma subsampling was not applied on the reference, since it alters the data.

A total of five compression schemes were evaluated. Each compression scheme was given a label, as stated before, for easier identification. A summary of the compression schemes can be found in Table 7.1. It should be noted that the QPs were selected to match the bitrates described above.

7.2.2 Objective quality evaluation

To analyze the performance of evaluated coding schemes, PSNR was selected as a full reference metric. PSNR values were computed with respect to the uncompressed reference. The computation is thus performed on the 4D light field after color and gamma corrections (point B in Figure 7.1).

The PSNR metric was adapted to better suit properties of light field images, and was computed analogously to what has been described in Chapter 6.1. We report it here for completeness. The PSNR value is computed on the Y channel as follows:

$$PSNR_Y(k,l) = 10\log_{10}\frac{255^2}{MSE(k,l)}, \tag{7.1}$$

in which k and l are the indexes of the acquired views. The $MSE(k,l)$ for each image is computed as follows:

$$MSE(k,l) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j) - R(i,j)]^2, \tag{7.2}$$

where m and n are the dimensions of one viewpoint image (i.e., $n = 625$, $m = 434$). $I(i,j)$ is the Y value for the selected acquired view in the evaluated 4D light field, whereas $R(i,j)$ is the corresponding value in the reference 4D light field. In the same way, the PSNR for the other two channels U and V is obtained. A weighted average [Ohm et al., 2012] is then computed as follows:

$$PSNR_{YUV}(k,l) = \frac{6PSNR_Y(k,l) + PSNR_U(k,l) + PSNR_V(k,l)}{8} \tag{7.3}$$

The mean of all viewpoint images is subsequentially computed to have an average value for PSNR for Y channel and for YUV:

$$PSNR_{X_{mean}} = \frac{1}{(K-2)(L-2)}\sum_{k=2}^{K-1}\sum_{l=2}^{L-1}PSNR_X(k,l), \tag{7.4}$$

in which $K = 15$ and $L = 15$ represent the number of perspective views, and $X = Y$ and $X = YUV$ for Y channel and for YUV channels, respectively.

7.2.3 Subjective assessment

The interactive and passive methodologies employed in this test were extensively described and compared in Chapter 3, to which we refer interested readers for more information on the test environment and planning.

Data analysis

Outlier detection was performed according to the guidelines defined in ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. One outlier was detected in both interactive and passive tests, and the relative scores were discarded, thus leading to 23 and 28 scores per stimulus, respectively. The MOS was computed, separately for each methodology, for each coding condition j (i.e., each content, codec and compression ratio) as follows:

$$MOS_j = \frac{1}{N}\sum_{i=1}^{N} m_{ij}, \tag{7.5}$$

where N is the number of participants and m_{ij} is the score for stimulus j by participant i. The corresponding 95% CIs were computed. To determine whether the results yield statistical significance, a one-sided Welch's test at 5% significance level was performed on the scores, with the following hypotheses:

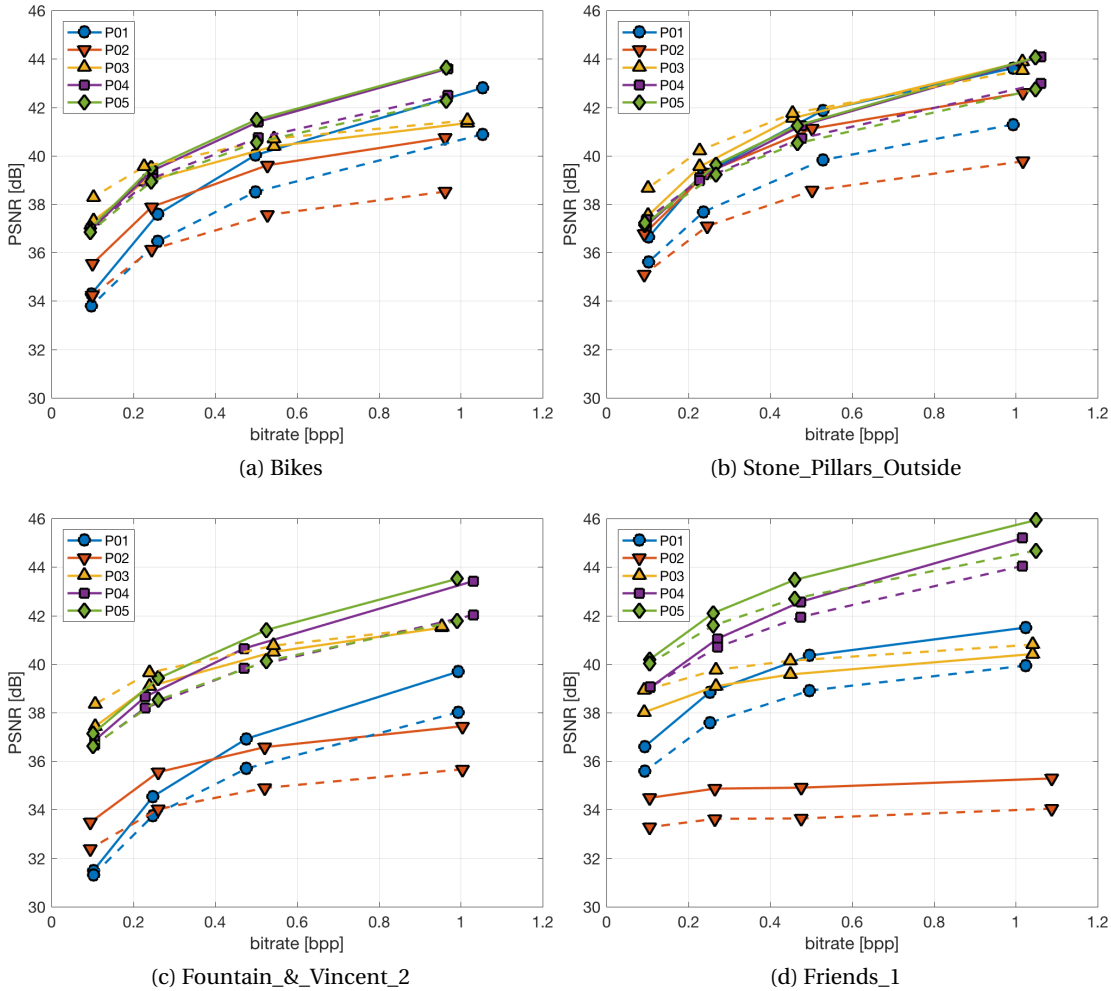
$$H0 : MOS_A \leq MOS_B$$

$$H1 : MOS_A > MOS_B,$$

in which A and B are the codecs that are being compared. The test was performed for each compression ratio and for each content, separately for each methodology. If the null hypothesis were to be rejected, then it could be concluded that codec A performed better than codec B for the given content and compression ratio at a 5% significance level.

7.3 Results and discussion

In this section, results of the objective and subjective quality assessments are presented. Results on the coding approaches presented in Section 7.1 will be discussed separately. First, the lenslet image compression is analyzed. Then, the 4D light field compression is discussed. The hybrid approach is compared to the other approaches. Finally, a comprehensive review of all the codecs is performed.



(a) Bikes

(b) Stone_Pillars_Outside

(c) Fountain_&_Vincent_2

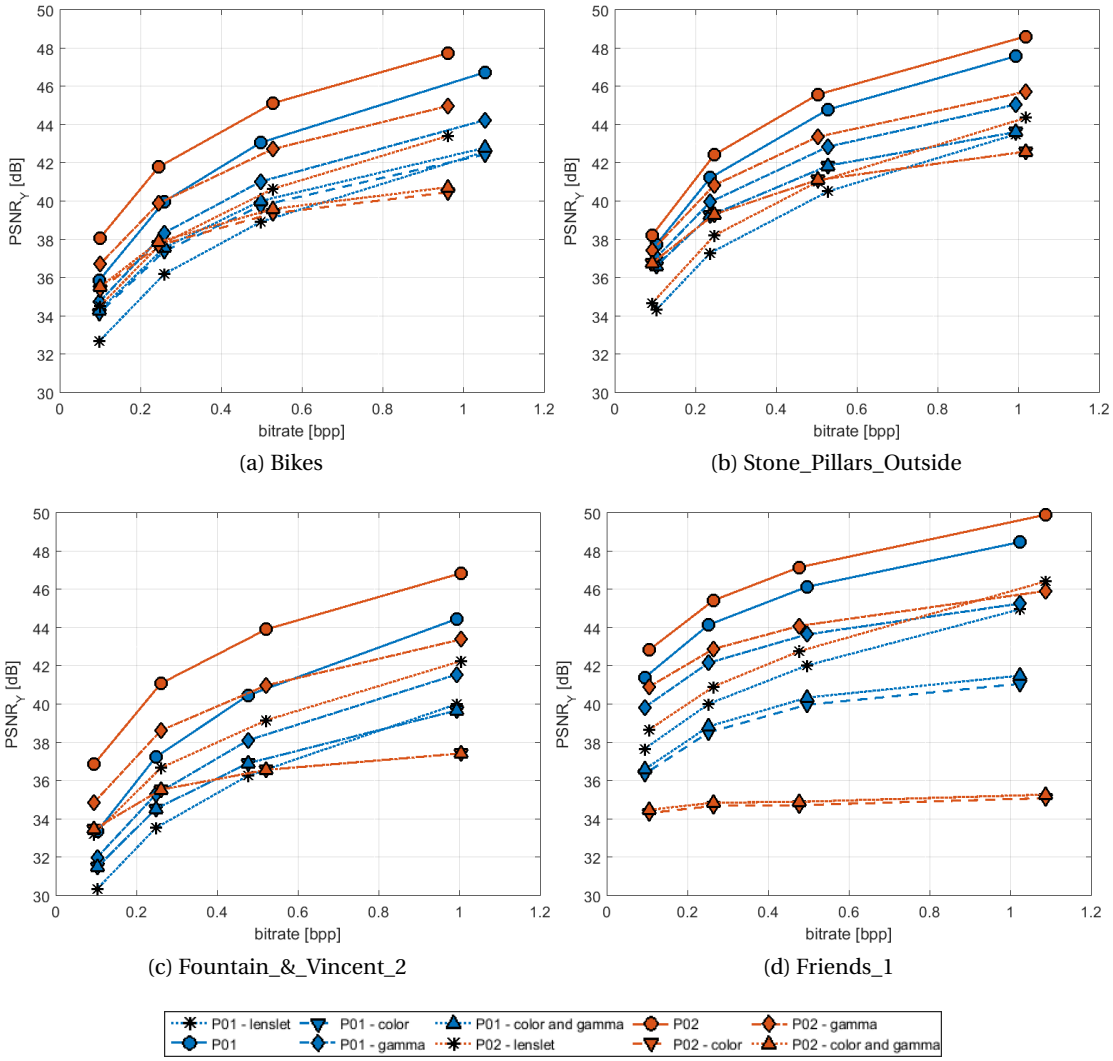
(d) Friends_1

Figure 7.7 – Rate distortion plots for Y channel (solid line) and for YUV channels (dashed line). PSNR was computed on the 4D light field after color and gamma correction. ©2017 IEEE

7.3.1 Compression of lenslet images

For PSNR computed on Y channel (Fig. 7.7, solid lines), the performance of the two codecs examined here (*P*01 and *P*02) strongly depends on the content, as it is common when computing PSNR. In general, *P*01 outperforms codec *P*02 for high bitrates. For low bitrates, *P*02 outperforms *P*01 for contents *Bikes* and *Fountain_&_Vincent_2*, and is outperformed in the remaining cases. PSNR computed on YUV channels (Fig. 7.7, dashed lines) shows similar trends.

Codec *P*02 has a particularly poor performance with content *Friends_1*, and in general performs worse than codec *P*01 for high bitrates. Results are particularly surprising since the codec proposed in *P*02 is supposed to improve the performance of HEVC Intra (anchor *P*01) with new prediction schemes. To better investigate the reasons behind this behaviour, we



(a) Bikes

(b) Stone_Pillars_Outside

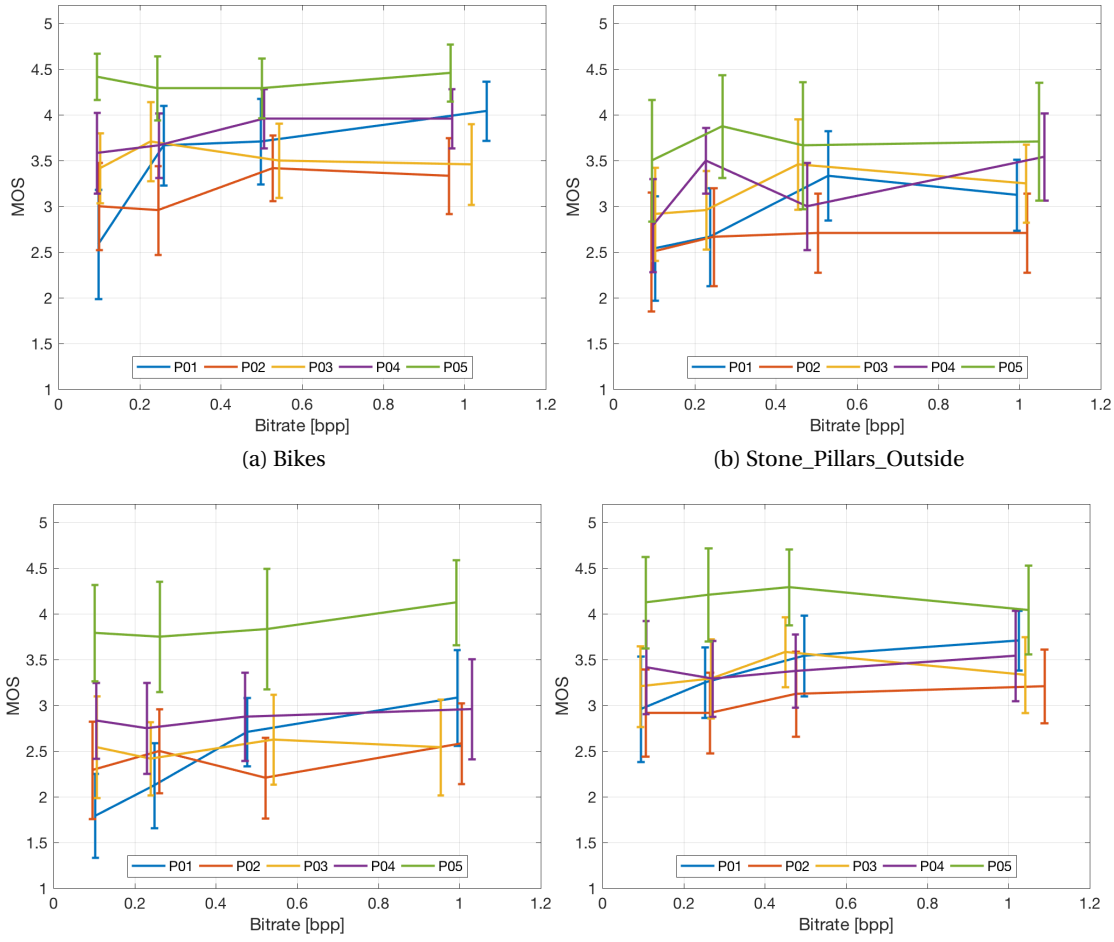
(c) Fountain_&_Vincent_2

(d) Friends_1

Figure 7.8 – Rate distorsion plots for Y channel. PSNR was computed at various stage of the pipeline (See Fig. 7.2). ©2017 IEEE

computed PSNR at different stages of the pipeline. Results from PSNR computation are shown in Figure 7.8. In particular, we computed PSNR on the 4D light field without any color or gamma correction, on the color-corrected 4D light field, on the gamma-corrected 4D light field and when both corrections were applied on the 4D light field. Additionally, we compute PSNR on the lenslet image prior to the transformation, to better assess the performance of the two codecs on 2D images. The PSNR was computed with respect to the uncompressed reference at the same stage of the pipeline.

Results show that, on the lenslet image and on the 4D light field without any correction, $P02$ always outperforms $P01$. On the gamma-corrected 4D light field, $P02$ performs better than $P01$ on half of the contents. When color correction is applied on the 4D light field, however, we



(a) Bikes

(b) Stone_Pillars_Outside

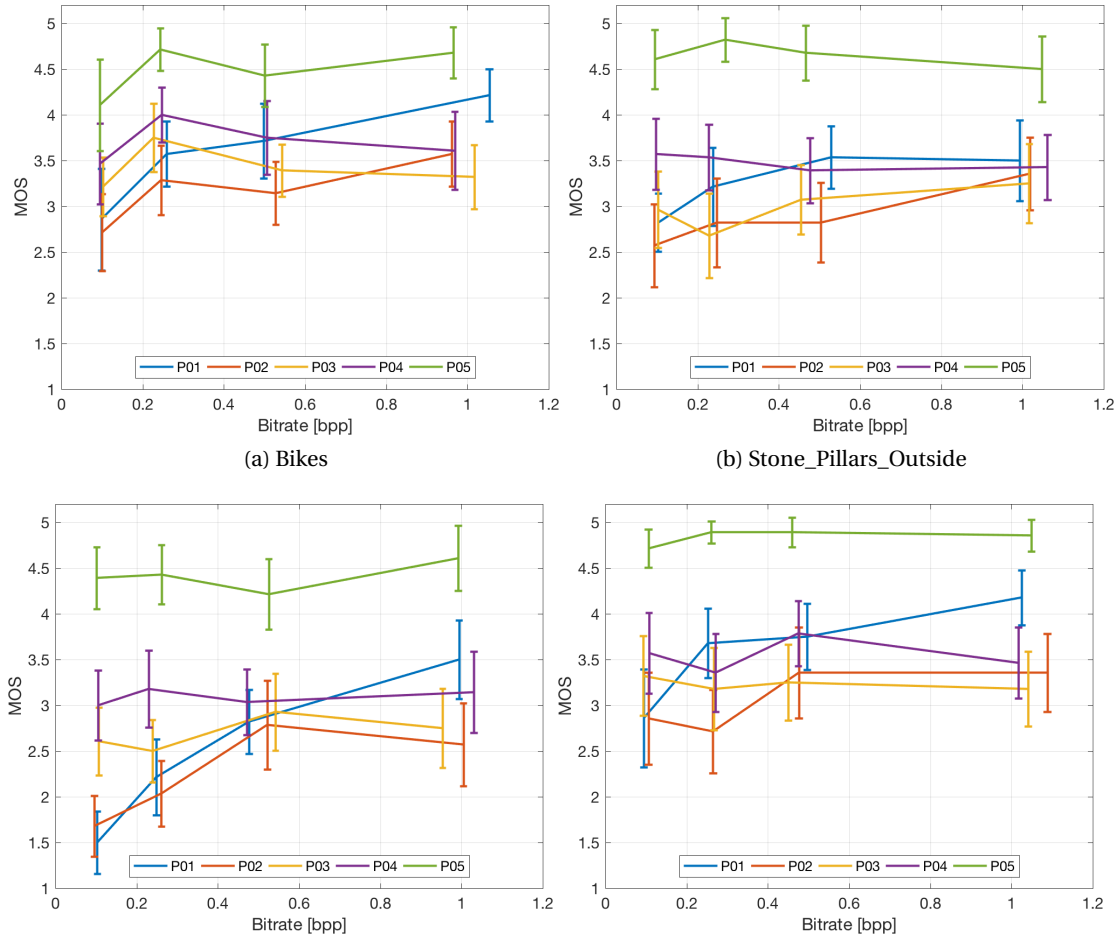
(c) Fountain_&_Vincent_2

(d) Friends_1

Figure 7.9 – Results of interactive subjective tests. MOS vs bitrate for all contents, with respective CIs. ©2017 IEEE

see a degradation in performance, with $P01$ outperforming $P02$ for high bitrates. This suggests that the prediction method, while working efficiently on compression of regular images, as proven by the results obtained on the lenslet image prior to transformation, adapts rather poorly to the peculiarities of light field images, and is more susceptible to errors after color correction is applied. Results from PSNR computed on YUV channels follow the same trend.

Results from both interactive and passive subjective evaluations show that $P01$ is performing significantly better than $P02$ for the highest bitrate. In particular, in the interactive test $P01$ is significantly better than $P02$ for all contents, whereas in the passive test it is significantly better for 3 out of 4 contents. For bitrate = 0.5 bpp, interactive tests show that $P01$ performs better than $P02$ for only 1 out of 4 contents, whereas passive tests indicate that it outperforms $P02$ on half of the contents. For lower bitrates (0.2 and 0.1 bpp) the difference between the two codecs is negligible.



(a) Bikes

(b) Stone_Pillars_Outside

(c) Fountain_&_Vincent_2

(d) Friends_1

Figure 7.10 – Results of passive subjective tests. MOS vs bitrate for all contents, with respective CIs. ©2017 IEEE

7.3.2 Compression of 4D light field

As discussed in section 7.1, we want to analyse the effect of downsampling the lenslet image prior to transformation to 4D light field. For this reason, we compare the performance of *P*04, which uses a chroma subsampled version of the lenslet image, with *P*05, which creates the 4D light field from the lenslet image which has not been subsampled (see Fig. 7.3).

For PSNR computed on Y channel (Fig. 7.7, solid lines), the two codecs have similar performance for all bitrates for contents *Bikes* and *Stones_Pillars_Outside*, whereas for contents *Fountain_&_Vincent_2*, *P*05 performs better than *P*04 for all bitrates. Although downsampling of chroma channels should not affect the Y channel, color correction is applied on RGB channels of the single views, which are then converted to YUV to compute the PSNR. The downsampling thus affects the Y channel as well.

Results are similar for PSNR computed on YUV channels (Fig. 7.7, dashed lines), although for content *Fountain_&_Vincent_2*, the difference between *P*04 and *P*05 is now negligible.

Results from subjective evaluations and pairwise comparison (Fig. 7.9, 7.10, 7.11 and 7.12) show a stronger preference for codec *P*05 when compared to codec *P*04. In particular, results from the interactive tests show that, for the highest bitrate, *P*05 performs significantly better than *P*04 for two out of four contents. For bitrate = 0.25 bpp, *P*05 performs better on three out of four contents, whereas for the remaining two bitrates (0.5 bpp, 0.1 bpp) it always performs significantly better than *P*04.On the other hand, results from the passive tests show that *P*05 always performs significantly better than *P*04, for all bitrates.

7.3.3 Hybrid compression of lenslet images

As seen in section 7.1, the third compression scheme *P*03 is compressing 4D light field and then converting them back to lenslet images. It is thus worthy of note to compare the performances of *P*01, *P*02 and *P*03, since they have the same input and output in the compression and decompression steps, although they use different approaches.

From the objective quality metric point of view, *P*01 and *P*03 outperform codec *P*02 for high bitrates. For low bitrates, *P*03 always outperforms *P*01 and *P*02, although in case of content *Stone_Pillars_Outside*, the difference between the codecs is negligible. For PSNR computed on YUV channels (Fig. 7.7, dashed lines), codec *P*03 outperforms the others for all contents. Interestingly enough, for codec *P*03 PSNR computed on YUV channels always has higher values than PSNR computed on the Y channel, while for all the other codecs the opposite is true. One possible explanation for this peculiar behavior is that the inverse color and gamma transformation applied before transforming the 4D light field back to lenslet has an effect on the final color performance, leading to better results in the YUV channels.

The subjective evaluation results do not show the same trends as the objective results (Fig. 7.9, 7.10, 7.11 and 7.12). In particular, results from the interactive tests show that for the lowest bitrate (0.1 bpp) *P*03 outperforms *P*01 on two out of four contents and never outperforms *P*02, whereas the passive tests show that *P*03 performs better than *P*01 on only 1 out of 4 contents, and performs better than *P*02 on 2 out of 4 (Fig. 7.12 (a)). For intermediate bitrates (0.5 bpp and 0.25 bpp), interactive tests show that *P*01 and *P*03 both perform significantly better than *P*02 on one out of four contents, whereas passive tests additionally show that *P*01 performs significantly better than *P*03 on half of the contents for both bitrates. For the highest bitrate, *P*01 performs significantly better than *P*03 on at least half of the contents (3 out of 4 in case of passive tests, 2 out of 4 in case of interactive tests), and outperforms *P*02 in the majority of cases (3 out of 4 in case of passive tests, 4 out of 4 in case of interactive tests).

For the objective quality evaluation, the hybrid scheme *P*03 performs better than the other lenslet compression schemes. However, results from the subjective evaluation suggest that the difference in performance with respect to *P*01 (simple HEVC Intra) is negligible for low

bitrates, and leads to poorer results for the highest bitrates.

Since *P*03 compresses lenslet images through transformation to 4D light field, it is useful to compare its performance to the performance of *P*04. For PSNR computed on Y channel (Fig. 7.7, solid lines), the performance of codecs *P*03 and *P*04 strongly depends on the content, as expected. For high bitrates, *P*04 performs better than *P*03, with the notable exception of content *Stones_Pillars_Outside*, in which codec *P*03 performs slightly better for all bitrates. For low bitrates, however, *P*04 performs slightly worse than *P*03 for all contents except *Friends_1*, for which *P*04 performs better than *P*03. PSNR computed on YUV channels (Fig. 7.7, dashed lines) show similar trends.

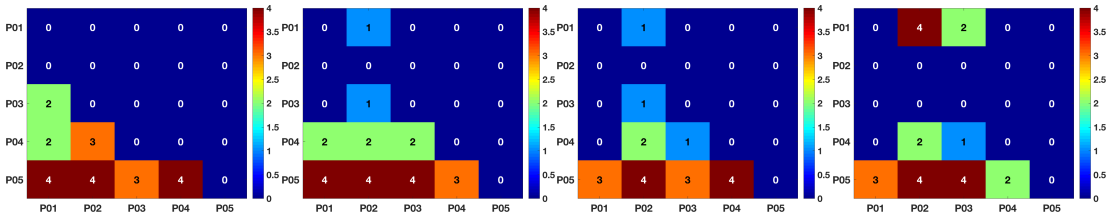
Results from subjective evaluation and pairwise comparison (Fig. 7.9, 7.10, 7.11 and 7.12), however, show that codec *P*03 never performs significantly better than codec *P*04. In particular, for the lowest bitrate, results from the interactive tests indicate that no codec performs significantly better than the other, whereas results from passive tests suggest that *P*04 performs better than *P*03 on 1 out of 4 contents. For bitrate = 0.25 bpp and 0.5 both interactive and passive tests agree that *P*04 outperforms *P*03 for 2 out of 4 contents and 1 out of 4 contents, respectively. For the highest bitrate, interactive tests indicate that *P*04 performs significantly better than *P*03 on 1 out of 4 contents, whereas for the passive tests they are statistically equivalent for all contents.

7.3.4 General discussion

In general, both objective and subjective results show that coding 4D light field (point B in Figure 7.1) leads to better performance when compared to coding lenslet images directly. In particular, pseudo-temporal ordering of 4D light field, obtained from RGB 444 lenslet image, performs significantly better than the other proposals for at least half of the contents for all bitrates examined in the subjective assessment of quality, showing that chroma subsampling of lenslet images can lead to a considerable reduction in visual quality. It is worth noting, however, that results from passive tests show that *P*01 performs statistically better than *P*04 on 2 out of 4 contents for the highest bitrate (Fig. 7.12 (d)).

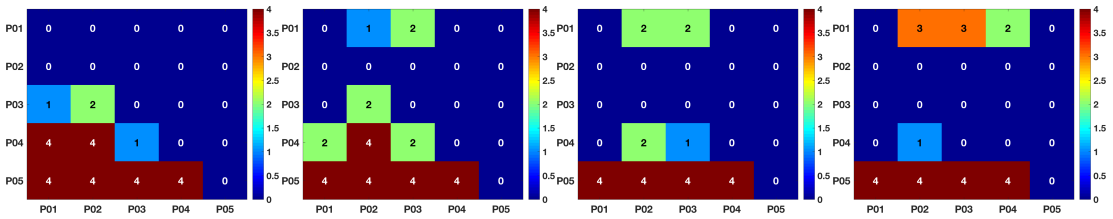
Comparison of different lenslet image compression algorithms shows that improvements in performance for 2D image coding do not necessarily result in better visual quality of light field image. In particular, whereas HEVC intra with LLE and SS has better performance in objective quality evaluation carried out on lenslet images and 4D light field without color correction, it performs significantly worse when color correction is applied. Further work on lenslet image compression should address the effect of color correction on the final 4D light field, and propose new strategies to appropriately cope with this issue.

Coding 4D light field has the benefit of not requiring any metadata to be correctly displayed. Moreover, it can be used for compression of contents acquired with different acquisition technologies. Since the most distorted views in the 4D light field can be discarded in the



(a) R4 (b) R3 (c) R2 (d) R1

Figure 7.11 – Pairwise comparison results for interactive subjective tests. Each cell contains the number of contents for which the null hypothesis was rejected, for each compression ratio. The null hypothesis is defined as $MOS_i \leq MOS_j$, in which i indicates the row and j the column of the matrix. ©2017 IEEE



(a) R4 (b) R3 (c) R2 (d) R1

Figure 7.12 – Pairwise comparison results for passive subjective tests. Each cell contains the number of contents for which the null hypothesis was rejected, for each compression ratio. The null hypothesis is defined as $MOS_i \leq MOS_j$, in which i indicates the row and j the column of the matrix. ©2017 IEEE

compression process, it also allows for bitrate saving. As we previously mentioned, the transformation to 4D light field on the decoder side is an additional step that would be not suitable for low-memory devices. Thus, if consumers' market is the desired target, a solution that does not require any transformation would be preferrable. In this case, coding 4D light fields seems the most suitable choice.

The additional step of converting to 4D light field on the decoder side is not an issue if the target is the professional market. However, fidelity to acquisition parameters has higher importance. As seen before, chroma subsampling leads to poorer performances, especially after color correction has been applied. On the other hand, coding the 4D light field data structure could lead to discarding metadata, which could be used in post-processing softwares, as well as potentially rejecting some heavily distorted views. In this case, both approaches presented in this chapter do not seem suitable. A new approach should be designed, aimed at high fidelity to acquisition parameters.

7.4 Conclusion

In this chapter, we presented two different coding approaches for light field image compression, based on the information to be encoded. The approaches were evaluated using both objective quality metrics and subjective experiments. Considering different predefined use cases, experimental results provide some insights on the impact of compression algorithms on the perceived quality. This reveals the necessity for further investigations and improvements of compression algorithms especially in terms of processing of the metadata related to light field rendering.

Our contributions can be outlined as such:

- We define two coding approaches for light field image compression, according to a generic pipeline for acquisition, transmission and rendering of light field contents. One approach takes as input the raw data after basic pre-processing, whereas the other requires the transformation to acquisition-independent 4D light field data structure prior to coding.
- We selected five different compression algorithms that adhere to either one of the aforementioned coding approaches. Specifically, two of the algorithms perform the compression on the raw data acquired by lenslet cameras, other two use video-based procedures to encode the perspective views composing the 4D data structure, and another employs an internal transformation to 4D light field to efficiently encode the raw data.
- We evaluated the compression algorithms on four bitrates, using 4 lenslet-based light field contents. The evaluation was performed using both objective quality metrics and subjective methodologies.
- By analysing the results, we showed that traditional perceptual-based encoding strategies, such as chroma subsampling, are not suitable for the compression of raw data, as they can lead to some non-intended effects on the final visual quality after the transformation to 4D light field.
- We also show that one coding approach, namely, compressing 4D light field data, yields significantly better results in terms of visual quality for all bitrates when compared to compressing lenslet images.

Regarding the use cases we defined at the beginning of the chapter, the 4D light field coding approach is particularly suitable for a general consumer scenario, since it does not involve additional computations at the decoder side to be properly rendered. Moreover, the coding approach does not require metadata to be successfully decoded and displayed, thus reducing the bitrate. Finding a successfull approach for the professional audience, however, is still an open issue. A new method for compressing lenslet images while taking into account color

fidelity must be designed for this type of market. Further research should focus on how to modify the proposed compression algorithm for light field images to further improve the performance and to meet the needs of all use cases.

8 Rendering-dependent encoding for light field tensor displays

In the previous chapters, we have observed and analyzed several approaches for light field compression that aim at improving coding efficiency. One of the main goals in designing a new compression solution, is to be able to reduce the amount of data that needs to be stored and trasmitted, while retaining an acceptable perceptual quality. Thus, taking into account the visualization scenario for the contents under compression is of extreme importance when deciding on which tools to use to perform the encoding.

Considering the possibilities that light field imaging offers in terms of rendering, optimizing a compression architecture in terms of visual quality can be proven challenging. The problem of assessing the effect of compression distortions on the rendered quality has been tackled in the past for image-based rendering. For example, Rizkallah et al. investigate the impact of compression of light field contents on extended focus and refocusing applications [Rizkallah et al., 2016]. Similarly, Perra et al. report the results of applying HEVC-based compression on light field refocusing [Perra and Giusto, 2018]. Adhikarla et al. analyse the effect of various distortions, including compression artifacts, on the visual quality of light field on 3D displays with motion parallax [Adhikarla et al., 2017].

Most of the compression solutions we have presented in previous chapters were designed and optimized for image-based rendering; in fact, the visual quality was conventionally assessed by measuring the level of distortion of the perspective views composing the 4D light field structure (see Chapter 6 and 7), even when different rendering procedures were evaluated (e.g., [Graziosi et al., 2014] for depth-based rendering). Nevertheless, the visual quality of light field contents under compression distortions may considerably vary under different types of rendering technologies. Moreover, rendering-agnostic solutions for light field compression may not be the most efficient solutions when specific displays are adopted. For example, multi-layer displays offer a multi-viewing experience while only rendering a few light-attenuating layers. Consequently, a large number of viewing angles can be experienced through the use of few, carefully created layer patterns. It is thus worth exploring whether new compression strategies can be specifically designed for multi-layer rendering, and how different approaches can affect the visual quality of the final rendered content.

(a) Back layer (b) Middle layer (c) Front layer

Figure 8.1 – Depiction of layer patterns generated from the light field perspective views.

In Chapter 5 we have presented a comparison among different methodologies and display solutions for visual quality assessment of light field contents on multi-layer displays. For this purpose, we introduced three coding strategies that could take advantage of the compressive capabilities of this type of display. However, as our analysis was mainly devoted to correlate the results obtained with different test conditions, we did not analyze which coding strategy could lead to the most efficient results. With this chapter, we aim at filling the gap by providing a thorough study of the effects of employing different approaches on the visual quality of light field contents in a multi-layer renderer.

8.1 Rendering-dependent coding strategies

Considering the peculiarities of multi-layer rendering, three viable alternatives for light field compression are envisioned and defined:

- **Compression of light field perspective views.** The first, rendering-agnostic solution performs the compression on the perspective views, which will then be transmitted through the communication channel. The layer patterns are created at the receiver side, after the compressed views have been decoded. This solution has the advantage of being adaptable to any rendering system, as no assumption on how the views will be rendered is made on the compression stage. However, as pointed out in [Takahashi et al., 2018], a large number of perspective views is needed to create a few light-attenuating layers. Thus, it might not be the most efficient solution when multi-layer displays are involved.
- **Compression of layer patterns.** The second solution performs the compression on the layer patterns directly, which can then be transmitted and rendered at the receiver side. This approach has the obvious advantage of compressing and transmitting only a few light-attenuating layers, thus gaining in compression efficiency. However, apart from being a rendering-dependent solution, which would require multiple transmissions for different devices, this approach might require ad-hoc algorithms to operate efficiently on the synthetic layer patterns, which are remarkably different from natural scenes (see Figure 8.1). Traditional image and video compression standards such as JPEG or HEVC

are optimized for natural scenes; thus, compressing the light-attenuating layer patterns may result in a sub-par performance.

- **Compression of focal stack.** It was shown in [Takahashi et al., 2018] that, in order to reconstruct the layer pattern from the focal stack, we only need a number of focused images equal to the number of layer patterns which need to be rendered. The focal stack can be easily compressed using conventional image and video standards, unlike layer patterns, and its limited amount of images should guarantee a better coding efficiency with respect to the first approach. However, the construction of layer patterns from focal stacks is riddled with additional errors, due to the approximation in the tensor factorization.

8.2 Experiment design

8.2.1 Dataset and coding conditions

Five light field contents were selected from a publicly available database [Řeřábek and Ebrahimi, 2016]. The contents were acquired with a Lytro Illum camera and processed using the Light Field Matlab Toolbox [Dansereau et al., 2013][Dansereau et al., 2015] to obtain a stack of 15×15 perspective image, each having a resolution of 625×434 pixels. Color and gamma corrections were applied on each perspective image for the rendering. To avoid unwanted distortions caused by the lenslet structure of the Lytro Illum camera, only the 9×9 central perspective views were selected for the test. The central perspective view from each content is displayed in Figure 5.1.

The three coding strategies described in Section 8.1 are employed for the test. For all three solutions, the state-of-the-art video encoding standard HEVC was employed for the compression, to ensure a fair comparison. To perform the encoding, the reference software HM was used [ITU-T Q.6/SG 16 and ISO/IEC JTC 1/SC 29/WG 11].

The layer patterns were created using the software implementation in [Takahashi, 2018]. To create the focal stack, the Light Field Matlab Toolbox was employed [Dansereau et al., 2013][Dansereau et al., 2015]. In our validating test, the number of layers was fixed to $L = 3$.

The compression solutions were evaluated at four bit-rates, namely $R1 = 537$ kB, $R2 = 134$ kB, $R3 = 67$ kB, and $R4 = 27$ kB, corresponding to 0.2, 0.05, 0.025 and 0.01 bpp, respectively. The bpp are computed with respect to the original size of the 9×9 perspective views. The bit-rates were carefully chosen to cover the visual quality space while providing reasonable and fair comparison among the listed compression solutions.

8.2.2 Objective quality evaluation

To evaluate the impact of the distorsions caused by the proposed algorithms, PSNR and SSIM were selected from the literature to objectively assess the visual quality of the contents. The layer patterns obtained from the uncompressed light field were used as reference for each content.

The metrics were applied separately to the luma channel Y and for each layer pattern image, as follows:

$$PSNR_Y(l) = 10\log_{10}\frac{255^2}{MSE(l)}, \tag{8.1}$$

$$SSIM_Y(l) = \frac{(2\mu_I\mu_R + c_1)(2\sigma_{IR} + c_2)}{(\mu_I^2 + \mu_R^2 + c_1)(\sigma_I^2 + \sigma_R^2 + c_2)}, \tag{8.2}$$

in which l is the index of each layer pattern used in the rendering, $MSE(l)$ is the mean square error, μ_I and μ_R are the mean values, σ_I^2 and σ_R^2 are the variances, and σ_{IR} is the covariance of the two perspective views in channel Y. Please note that unlike previous chapters, we are now computing the metrics on the layer patterns. PSNR was computed for chrominance channels U, V following Equation 8.1, and a weighted average [Ohm et al., 2012] was calculated as follows:

$$PSNR_{YUV}(l) = \frac{6PSNR_Y(l) + PSNR_U(l) + PSNR_V(l)}{8} \tag{8.3}$$

The average PSNR value for Y channel was then computed across the viewpoint images:

$$\widehat{PSNR}_Y = \frac{1}{L}\sum_{l=1}^{L} PSNR_Y(l), \tag{8.4}$$

in which $L = 3$ represent the number of layer patterns. $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ were analogously computed following Equation 8.4.

8.2.3 Subjective quality assessment

For our experiments, the DSIS with 5-point grading scale (*5-Imperceptible, 4-Perceptible but not annoying, 3-Slightly annoying, 2-Annoying, 1-Very annoying*) was selected, according to the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. For this test, we will analyse the results obtained with Variant B (see Chapter 5), which presents a user-driven intermittent presentation of the impaired and reference stimuli.

Participants were asked to rate the quality of the test stimuli when compared to the uncompressed reference. They could access the reference content by pressing a specific key, and they could return to the test content by pressing another designated key. Participants were only allowed to give a score when the test contents was being rendered on the screen, and after at least one full switch between test and reference stimulus. In order to accustom the participants with what distortions to expect in the test images, a training session was organized before the experiment. Three training samples, created by compressing one additional content on the test bit-rates, were manually selected by expert viewers.

All the compressed stimuli were shown in one session. Additionally, two hidden references per content were added to the test: one consisted in the layer patterns generated from the uncompressed stack of perspective views, while the other was created from the uncompressed focal stack. Thus, a total of 70 stimuli were evaluated. The display order of the stimuli was randomized for each participant, and the same content was never displayed twice in a row.

Two laboratory settings were used for our tests, in the facilities of the École Polytechnique Fédérale de Lausanne (EPFL) and Nagoya University (NU).

In EPFL, a laboratory for subjective video quality assessment, which was set up according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012], was used for the test. A 27-inch Apple Display with native resolution of 2560 × 1440 pixels was used. The monitor settings were adjusted according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level of 0.2 cd/m^2. The controlled lighting system in the room consisted of adjustable neon lamps with 6500 K color temperature against mid-grey background walls. The illumination level measured on the screens was 18 lux. Conforming to requirements in ITU-R Recommendation BT.2022 [ITU-R BT.2022, 2012], the distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content. However, subjects were allowed to move further or get closer to the screen. A total of 20 subjects (10 males and 10 females) participated in the tests, amounting to 20 scores per stimulus per variant. Subjects were between 18 and 35, with a mean age of 23.29 years old. Before starting the test, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively.

In NU, a controlled environment was selected to perform the experiment. However, no calibration on the lighting system for the room was conducted. A prototype multi-layer display was used to perform a pilot evaluation [Kobayashi et al., 2017]. A total of 11 subjects (all males)

took part in the test. Subjects were between 18 and 35, with a mean age of 23.28 years old. Before starting the test, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively.

8.3 Statistical analysis

Outlier detection and removal was performed on the results, independently for each test, according to the ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. No outlier was detected in either batch of scores. After outlier removal, the MOS was computed for each stimulus, independently for each methodology. The corresponding 95% CIs were computed assuming a Student's t-distribution.

In order to perform a benchmark of the objective quality metrics in predicting the subjective quality of light field contents rendered through the simulator, or the multi-layer display, several fittings were applied to the PSNR and SSIM values, following the ITU-T Recommendation P.1401 [ITU-T P.1401, 2012]. In particular, first order and third order fittings were used to compare the objective quality metrics to the MOS values. RMSE, PCC, SRCC and OR were computed for accuracy, linearity, monotonicity and consistency, respectively.

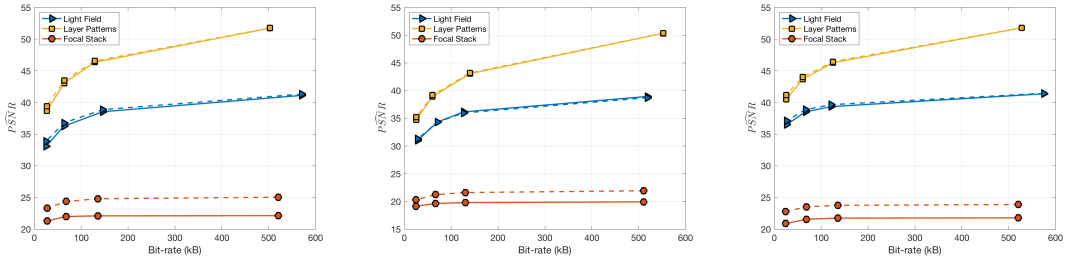
8.4 Results and discussion

In this section, results of objective and subjective evaluations are discussed, and a benchmark of objective quality metrics for visual quality prediction is given. First, in Section 8.4.1 we show the results of the objective evaluation, and we present and compare the results of the subjective assessment obtained with the two displays. Then, in Section 8.4.2 we carry out a benchmarking of the objective quality metrics using both subjective assessment tests as the ground truth.

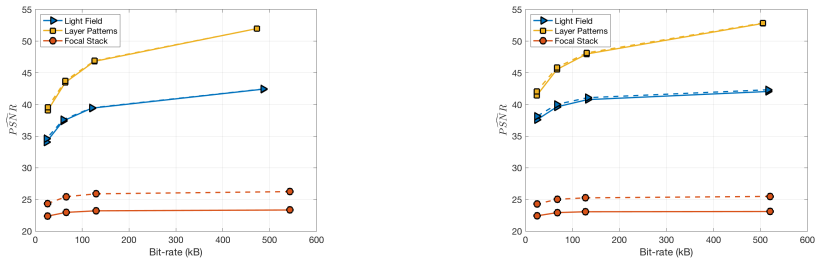
8.4.1 Objective and subjective results

Figures 8.2 and 8.3 show the results of the objective quality metrics, for all bitrates. Plots show that compressing the focal stack leads to a sharp decrease in performance when $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$ are used as metric, with a drop of ~ 30dB for high bitrates, and of ~ 15dB for low bitrates. Results for $\widehat{SSIM}_Y$ also indicate that the focal stack approach leads to reduced quality in the layer pattern data.

Figure 8.4 depicts the MOS scores obtained with the use of a simulator, whereas Figure 8.5 illustrates the MOS scores collected using a prototype multi-layer display. It has already been observed in Chapter 5 that the scores collected with different displays do not exhibit very good

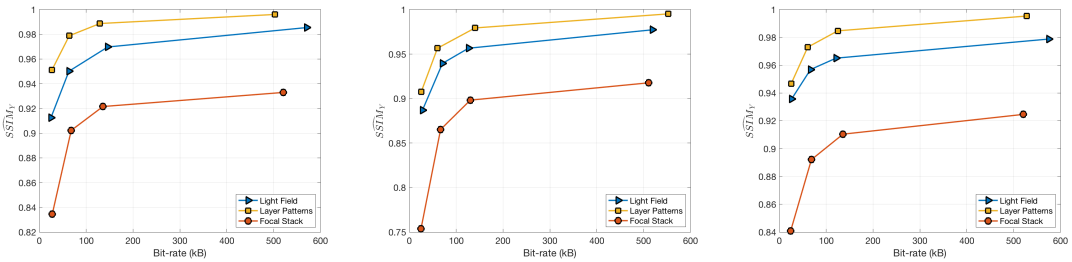


(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside

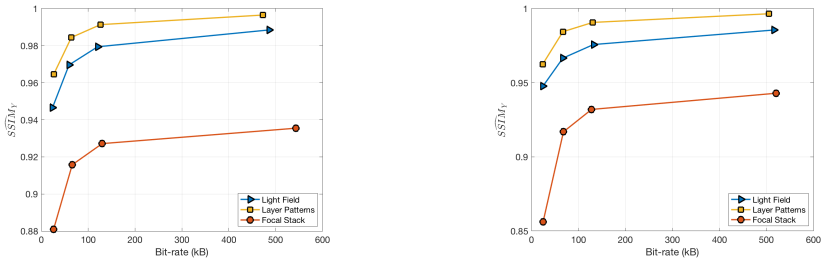


(d) Fountain_&_Vincent_2 (e) Friends_1

Figure 8.2 – Results of $\widehat{PSNR}_Y$ (solid line) and $\widehat{PSNR}_{YUV}$ (dashed line) vs bitrate for different contents.

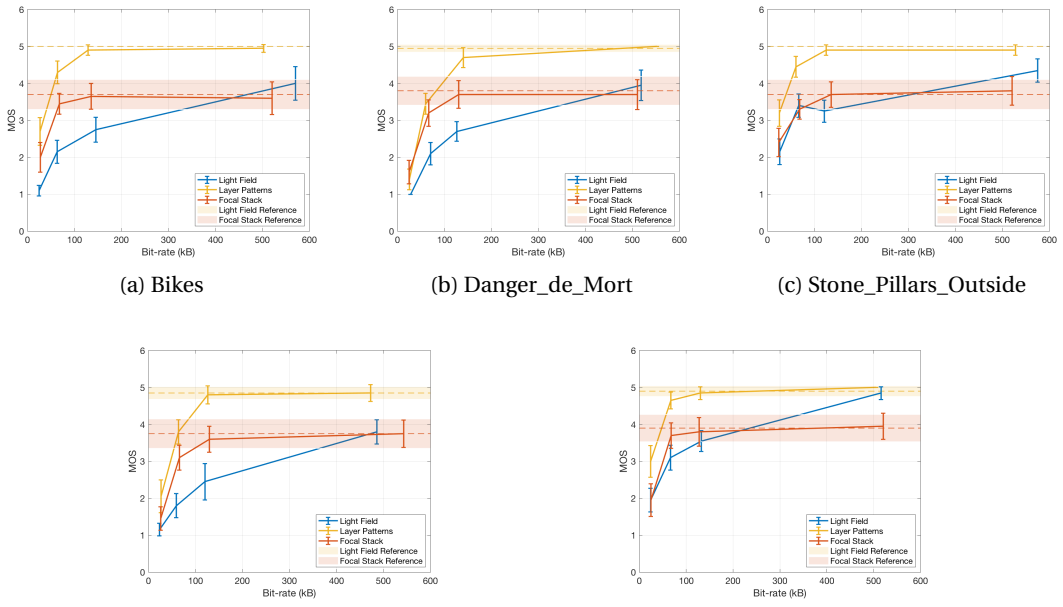


(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside



(d) Fountain_&_Vincent_2 (e) Friends_1

Figure 8.3 – Results of $\widehat{SSIM}_Y$ vs bitrate for different contents.



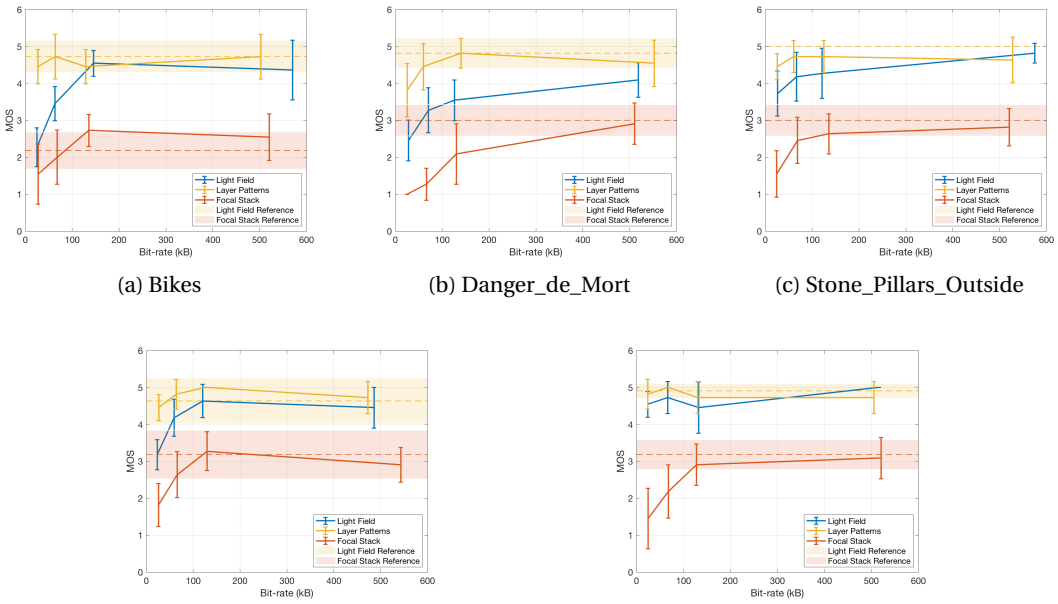
(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside

(d) Fountain_&_Vincent_2 (e) Friends_1

Figure 8.4 – MOS vs bitrate for different contents, with respective CIs. Results obtained using the simulator (see Annex C).

correlation (see Section 5.3.3). It is evident by considering how the first two approaches (compression of light field and layer patterns, respectively) consistently achieve near-transparent quality, regardless of the bitrate, when the multi-layer display is employed in the evaluation (Figure 8.5). On the other hand, compression of focal stack data leads to strongly perceived distortions, as the scores associated with the uncompressed reference are consistently under the MOS score of 4. Although the compression of focal stack data never reaches transparent quality when using the simulator (Figure 8.4), it performs comparably to a direct compression of light field data for high bitrates, and in some cases outperforms the latter.

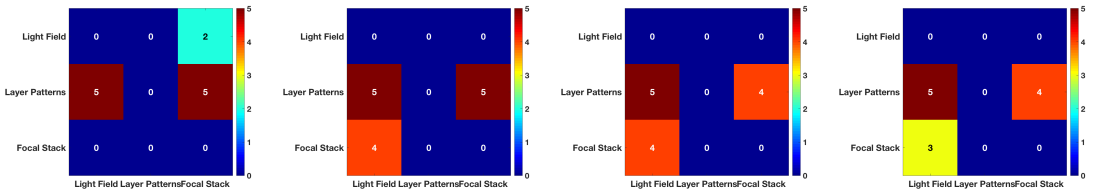
It is interesting to notice the difference in the CIs associated with the reference layer patters between the scores assigned using either display. When the simulator is adopted, the CIs are consistently small. This is to be expected, considering that the DSIS variant that is being used highlights subtle differences - or, in the case of the reference, the lack thereof. However, it is notable to see that the same does not happen when the multi-layer display is used. In this case, the large CIs would suggest that alterations and artifacts were perceived even when no difference between test and reference content was materially present. It would be worth exploring whether the uncertainty associated with the scores is a reflection of the hardware limitations of the prototype multi-layer display, which may interfere with the perception of distortions in subjective tests.



(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside

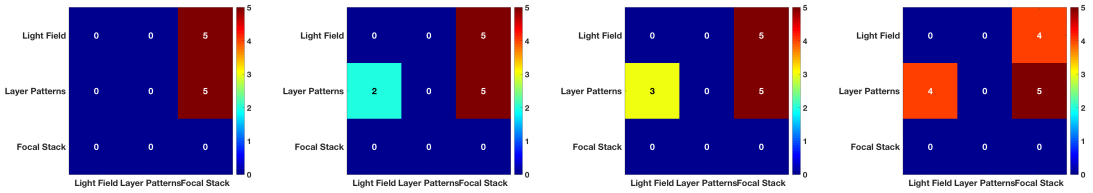
(d) Fountain_&_Vincent_2 (e) Friends_1

Figure 8.5 – MOS vs bitrate for different contents, with respective CIs. Results obtained using the prototype display.



(a) R1 (b) R2 (c) R3 (d) R4

Figure 8.6 – Pairwise comparison of codecs for different bitrates, for results obtained using the simulator.



(a) R1 (b) R2 (c) R3 (d) R4

Figure 8.7 – Pairwise comparison of codecs for different bitrates, for results obtained using the prototype display.

Table 8.1 – Performance indexes for the comparison among different objective quality metrics, fitted to the MOS scores obtained with the simulator.

	$[\widehat{PSNR}_Y, MOS]$				$[\widehat{PSNR}_{YUV}, MOS]$				$[\widehat{SSIM}_Y, MOS]$			
	PCC	**SRCC**	**RMSE**	**OR**	**PCC**	**SRCC**	**RMSE**	**OR**	**PCC**	**SRCC**	**RMSE**	**OR**
Linear fitting	0.4563	0.5899	1.0164	86.67%	0.4809	0.5912	1.0014	88.33%	0.5982	0.6833	0.9153	80.00%
Cubic fitting	0.6502	0.5899	0.8678	83.33%	0.6490	0.5912	0.8690	85.00%	0.6823	0.6833	0.8351	85.00%

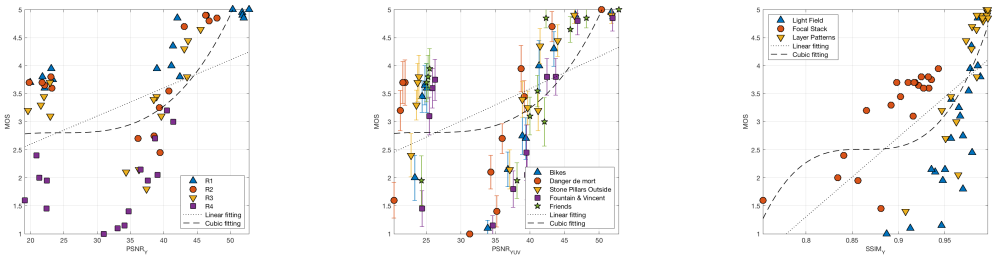
Table 8.2 – Performance indexes for the comparison among different objective quality metrics, fitted to the MOS scores obtained with the multi-layer display.

	$[\widehat{PSNR}_Y, MOS]$				$[\widehat{PSNR}_{YUV}, MOS]$				$[\widehat{SSIM}_Y, MOS]$			
	PCC	**SRCC**	**RMSE**	**OR**	**PCC**	**SRCC**	**RMSE**	**OR**	**PCC**	**SRCC**	**RMSE**	**OR**
Linear fitting	0.8999	0.9110	0.5075	40.00%	0.9031	0.9135	0.4998	40.00%	0.9101	0.8904	0.4821	31.67%
Cubic fitting	0.9191	0.9110	0.4584	33.33%	0.9285	0.9135	0.4322	26.67%	0.9348	0.8904	0.4134	23.33%

The graphs show that operating the compression on the layer patterns seems to be the preferable solution, as its performance is either statistically equivalent or better than the other solutions at all bitrates, for both displays. However, no definite answer can be given whether compressing the focal stack would be a preferrable solution with respect to encoding the entire light field, as contradicting results are obtained in the two tests. This is particularly evident when analyzing the results of the pairwise comparison among the scores, depicted in Figure 8.6 for the simulator test, and in Figure 8.7 for the prototype one. The boxes represent the number of contents for which the compression approach in each row is significantly better than the approach in each column. In the first case, encoding the layer patterns is the clear winning solution, as it outperforms the other two approaches in at least 4 out 5 contents; encoding the focal stack is the second preferred solution, as it fares better than encoding the light field data on more than half the contents for all bitrates, except the highest. The results, however, are overturned when considering the prototype multi-layer display (Figure 8.7): in this case, compression applied on the focal stack is never significantly better than the other two approaches, and it is nearly always outperformed. Layer pattern and light field encoding are statistically equivalent for high bitrates, but for lower bitrates the first approach leads to significantly better results.

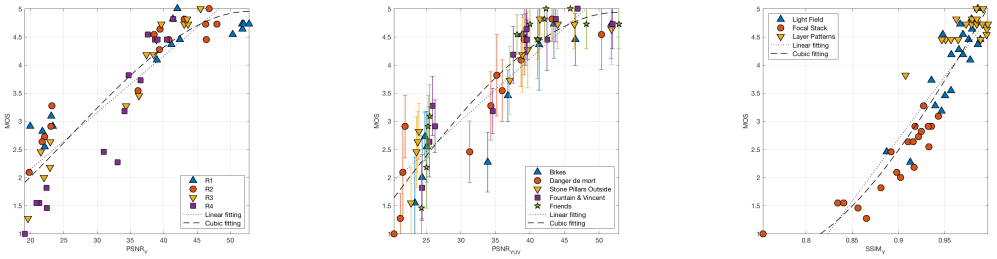
8.4.2 Benchmarking of objective quality metrics

Figures 8.8 and 8.9 depict the scatter plots between the objective quality metric results and the corresponding MOS scores, obtained using the simulator and the prototype multi-layer display, respectively. Tables 8.1 and 8.2 report the performance indexes for each metric, using linear and monotonic cubic fitting. Low values of PCC and SRCC confirm that the objective quality



(a) MOS as function of $\widehat{PSNR}_Y$. (b) MOS as function of $\widehat{PSNR}_{YUV}$. (c) MOS as function of $\widehat{SSIM}_Y$.

Figure 8.8 – Comparison of performance of different objective quality metrics in predicting the MOS scores obtained with the simulator, along with linear and cubic fittings. Points are differentiated by compression ratio (a), by content (b), and by compression solution (c).



(a) MOS as function of $\widehat{PSNR}_Y$. (b) MOS as function of $\widehat{PSNR}_{YUV}$. (c) MOS as function of $\widehat{SSIM}_Y$.

Figure 8.9 – Comparison of performance of different objective quality metrics in predicting the MOS scores obtained with the multi-layer display, along with linear and cubic fittings. Points are differentiated by compression ratio (a), by content (b), and by compression solution (c).

metrics are very poorly correlated with the subjective scores collected using the simulator. Among the three metrics, $\widehat{SSIM}_Y$ seems to perform slightly better (PCC = 0.6823 and SRCC = 0.6833 when cubic fitting is applied), although high levels of RMSE and OR indicate that accuracy and consistency are still lacking.

When considering the MOS scores collected using the multi-layer display, however, results show a strong correlation with all the objective quality metrics; again, $\widehat{SSIM}_Y$ is the best performing one, achieving PCC = 0.9348 and SRCC = 0.8904 with cubic fitting.

Results show that all objective quality metrics are good predictors for visual quality of light field contents when visualized through a prototype display. However, considering the level of uncertainty associated with the scores, as proven by the large CIs for the uncompressed reference stimuli, we are hesitant in recommending the use of said objective quality metrics as quality predictors for compression artifacts. As shown also in Chapter 5, the method employed to generate the layer patterns seems to have a higher impact on the final subjective scores, at least when the prototype display is used. Thus, it is safe to assume that objective

quality metrics would give a reliable estimation on which method for layer pattern generation would lead to the best visual quality. However, when compression artifacts need to be assessed and compared, new objective quality metrics should be developed. Moreover, particular care should be given in assessing whether the hardware limitations of the display in use restrict the perception of compression distortions; in that case, the use of an error-free, ideal rendering scenario should be preferred.

8.5 Conclusion

In this chapter, we presented objective and subjective quality assessment results of different coding strategies for multi-layer-rendered light field contents. For the subjective assessment, we performed the tests on both a multi-layer simulator and a prototype display, showing that the scoring distribution varies considerably between the two. Finally, we benchmarked the performance of objective quality metrics in predicting the visual quality of light field contents on both displays.

A summary of our contributions:

- We define three coding strategies for multi-layer-based rendering. The first strategy performs the compression on the perspective views forming the 4D light field structure, delegating the conversion to a multi-layer-appropriate format to the decoder side. The second strategy directly encodes the layer patterns that will be visualized in the multi-layer display. Finally, in the third strategy a stack of refocused images is encoded and transmitted, and the layer patterns are generated at the decoder side before the rendering. Possible advantages and drawbacks of all approaches are presented.

- We design an objective and subjective quality evaluation campaign to assess the performance of the aforementioned coding approaches. For the objective evaluation, common image metrics are applied on the layer patterns obtained from the encoded streams, whereas for the subjective evaluation, both a prototype multi-layer display and a simulator were employed.

- We perform a comparison of the coding strategies using the outcomes of the objective and subjective tests. Results show that applying the compression on the layer patterns leads to a superior performance with respect to the other two approaches.

- We show that the method employed for generating the layer patterns has great impact on the visual quality of the rendered contents. In particular, we observe that, by using the focal stack to create the layer patterns, transparent quality is never achieved, even in the absence of compression distortions.

- We observe that compression artifacts are not quite perceivable when the prototype display is used to render the contents. In particular, results of the subjective evaluation

performed on the prototype display show that a near-constant level of quality is often maintained across all bitrates, when the first or second strategy is adopted. Conversely, generating the layer patterns from focal stack leads to strongly perceived artifacts, even at high bitrates.

- We benchmark existing objective quality metrics with the subjective experiments we performed on the two visualization settings. Results show that poor correlation is achieved between the fitted objective measurements and the subjective scores obtained with the simulator. On the other hand, the objective quality metrics are in agreement with the results obtained with the multi-layer display, where the compression artifacts were not strongly perceived.

It appears that compression artifacts were not perceived when the prototype display was employed, as showed by the fact that no difference in quality was perceived across the bitrates for the first two strategies, on the near majority of the contents. It is yet to be determined whether these results are a consequence of the hardware limitations, which hinder the QoE associated with the display. Thus, our recommendation would be that it is not advisable to employ objective quality metrics to predict the visual quality of encoded light field contents in an ideal rendering scenario, as the simulator is reproducing. However, they may be suitable for predicting the quality associated with the method employed to generate the layer patterns, as they showed to be aligned with the results obtained with the prototype multi-layer display.

New objective quality metrics should be designed to successfully estimate the impact of coding solutions on the visual quality of light field contents. Moreover, great care should be employed in designing subjective tests to assess the visual quality associated with multi-layer displays. In particular, the hardware limitations linked to prototype displays may lead to a poor QoE, which can be reflected on the distribution of the collected subjective scores.

Part III

Towards new compression solutions for light field contents

9 Encoding disparity information for lenslet-based light field images using graph learning

Disclaimer: This chapter was adapted from the following article, with permission from all co-authors and publishing entities:

Irene Viola, Hermina Petric Maretic, Pascal Frossard, Touradj Ebrahimi, "A graph learning approach for light field image compression," Proc. SPIE 10752, Applications of Digital Image Processing XLI, 107520E (17 September 2018). DOI: https://doi.org/10.1117/12.2322827

Personal contribution: I formulated the problem and designed the compression algorithm, while the graph learning part was curated by the other main author. I performed the validating test and carried out the analysis of the results.

In the previous chapters we have stressed the importance of reducing the size of the acquired light field data to acceptable dimensions for transmission and storage. Notable gains can be obtained by exploiting the naturally occurring redundancies in the light field representation, in order to minimize the size of the data without compromising the perceptual visual quality.

The initiative launched by the JPEG standardization committee, JPEG Pleno, is indicative of the interest on finding a standard framework for efficient storage and delivery of plenoptic contents, including light fields, point clouds, and holograms. In particular, JPEG Pleno aims at finding the minimum number of representation models for these types of content, while offering, when necessary, interoperability with existing standards, such as legacy JPEG and JPEG 2000 formats. For the past years, the JPEG committee has been actively pursuing the definition of a new standard representation and compression algorithm for light field images. In 2017, a CFP for light field coding solutions was issued jointly with ICIP 2017 Grand Challenge on light field image coding (see Chapter 6). The majority of the collected solutions were optimized for

compressing light field images with narrow baselines and dense angular sampling, such as lenslet-based light field images. An overview of recent compression solutions for lenslet-based light field images is provided in Chapter 2.

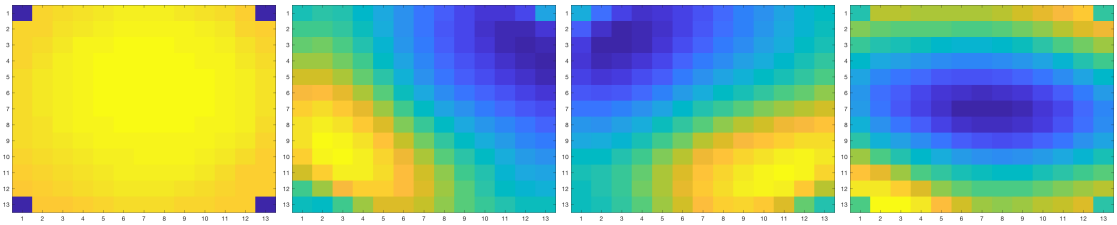
Among others, graph-based methods have been recetly proposed to efficiently compress light field data. Graph based methods for light field compression include the work of Maugey et al. [Maugey et al., 2013] where graph based representations are used to describe multiview geometry, and the work of Su et al. [Su et al., 2017], which adopts graph based representations to model colour and geometry. In their approach, the vertices of the graph correspond to each pixel in sub-aperture images, while the edges built from disparity information connect pairs of pixels across two images. In the work of Chao et al. [Chao et al., 2017], a graph between pixels is constructed through a Gaussian kernel, and use a graph lifting transform to compress light field images before demosaicking. However, the graph construction step can be proven to be delicate. Recently, graph learning methods have been attracting increasing interest for their capability to automatically infer the relationship among nodes. Among the first approaches, Dong et al. [Dong et al., 2016] consider a smooth signal model to infer the graph structure. Kalofolias [Kalofolias, 2016] explores a similar model, adding an option to promote graph connectivity, and offering a computationally more efficient solution. Fracastoro et al. [Fracastoro et al., 2016] propose a graph learning method for image coding, formulating a rate-distortion optimization problem that takes into account the cost of sending the graph.

In this work, we use the recent advances in graph learning to devise a new compression scheme for light field images that exploits the redundancy in the light field structure to reconstruct the entire 4D light field from an arbitrarily chosen subset of perspective views. Graph learning techniques are used to estimate the similarities among neighboring views. To reduce the impact of the graph on the overall data volume, in this approach the graph is constructed considering each view as a vertex. Edge weights relative to each pair of perspective images are learned from the data. The graph is then losslessly compressed and transmitted along with the selected perspective views. At the decoder side, an optimization problem is solved to optimally reconstruct the 4D light field. Results show the superiority of our method with respect to state-of-the-art solutions in light field compression.

9.1 Graph signal processing preliminaries

In this section, some basic notions in graph signal processing are introduced. A more detailed description can be found in [Shuman et al., 2013]. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ be an undirected, weighted graph with a set of m vertices $\mathcal{V}$, edges $\mathcal{E}$ and a weighted adjacency matrix W. Value W_{ij} equals 0 if no edge is present between i and j; otherwise, it designates the weight of that edge. The graph signal is defined as a function $y : \mathcal{V} \rightarrow \mathbb{R}$, where $y(v)$ denotes the signal value on a vertex v. The graph Laplacian is defined as

$$L = D - W, \tag{9.1}$$



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

Figure 9.1 – First 4 components of a PCA decomposition for the luminance component of *Bikes*. Each point represents one view.

where D is a diagonal matrix containing node degrees. As a real symmetric matrix, the graph Laplacian has a complete set of orthonormal eigenvectors $\chi = \{\chi_0, \chi_1, ..., \chi_{m-1}\}$ with a corresponding set of non-negative eigenvalues. Zero appears as an eigenvalue with multiplicity equal to the number of connected components of the graph, while the spectrum of the Laplacian matrix satisfies

$$\sigma(L) = \{0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{m-1}\}. \tag{9.2}$$

We can then define the graph Fourier transform $\hat{y}$ of a signal y at frequency λ_l as the expansion:

$$\hat{y}(\lambda_l) = \langle y, \chi_l \rangle = \sum_{i=1}^{m} y(i)\chi_l^*(i), \tag{9.3}$$

and the inverse graph Fourier transform as

$$y(i) = \sum_{l=0}^{m-1} \hat{y}(\lambda_l)\chi_l(i). \tag{9.4}$$

Here, the graph Laplacian eigenvectors form a Fourier basis; it straightforward to see that the corresponding eigenvalues carry a notion of frequency.

9.2 Proposed approach

In this work, we propose a novel approach for light field image coding using graph learning. We exploit the extensive similarities between the views by capturing them in a graph that models their relationship. We then select a subset of views to be compressed and transmitted along with the graph, which will be used to recover the remaining views. This approach allows for better compression quality in sampled views, as more bits can be allocated to encode them.

The intuition behind our approach stems from observing the presence of smoothness among

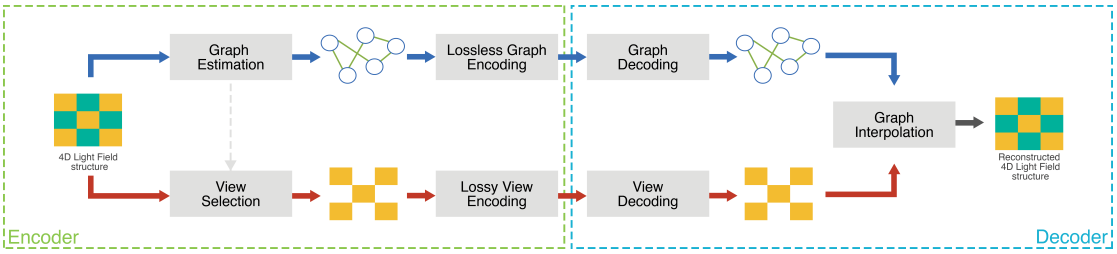


Figure 9.2 – Overview of the compression scheme.

neighboring images in a 4D light field structure. The idea is confirmed by simple PCA analysis of the signal, which shows a smooth, slowly transitioning behavior, further strengthening the suggestion of smoothness among neighboring views. Figure 9.1 shows this phenomena for *Bikes*, with each of the points in a PCA component representing one view. Graph signal processing is traditionally used to properly capture this smoothness on an irregular structure, defining notions equivalent to those seen in regular signal processing [Shuman et al., 2013].

In order to maximally exploit signal smoothness, we use a graph learning algorithm to obtain a structure our signal is most smooth on. We construct a graph with each vertex representing one view, and learn edges modelling relationships between corresponding views. This adaptive graph structure, as opposed to a simple fixed grid graph, ensures a much better signal representation at an acceptable cost. Graphs are convenient structures for compression, as they encode a large amount of information through their Fourier domain, while retaining sparsity in the vertex domain. In fact, since a graph Fourier domain is obtained through eigendecomposition of the graph Laplacian matrix, the vectors representing the Fourier basis are different for every graph. Even more, if the graph has been constructed in a way that ensures signal smoothness, the Fourier basis vectors will be representative of this specific set of signals, and the signal will be smooth in this basis.

9.2.1 Overview of the compression scheme

The general structure of our proposed compression scheme is depicted in Figure 9.2. The encoder first estimates the graph between all the views of our image, which is then losslessly encoded. At the same time, the encoder will also select a subset of views that are to be compressed directly, as opposed to the rest, which will be estimated from them. It then performs a lossy compression of the selected subset. The decoder receives an encoded graph and a subset of views. After decoding both, it solves an optimization problem to estimate the remaining views, and, in low bitrates, to improve the existing ones. A MATLAB implementation of the proposed solution can be found at the following link: https://github.com/mmspg/light-field-graph-codec.

9.2.2 Encoder

Graph estimation represents the most crucial step in our encoding scheme, as it models the dependencies among perspective views and allows for a faithful reconstruction of the non-encoded views. To obtain a graph that will best describe relationships among the 4D light field, while emphasizing signal smoothness on its structure, we resort to a graph learning technique. We consider each view as a vertex, to minimize the number of weights that need to be encoded, thus reducing the overhead created by sending them. As described in [Kalofolias, 2016], the following optimization problem yields a graph representing smooth signals, while promoting connectedness and providing a mean to control graph sparsity. The problem reads as follows:

$$\mathbf{argmin}_{W \in \mathcal{W}_m} \quad tr(Y^T L Y) - \alpha \mathbf{1}^T log(W\mathbf{1}) + \beta \|W\|_F^2 \tag{9.5}$$

$$L = D - W \tag{9.6}$$

$$\mathcal{W}_m = \{W \in \mathbb{R}_+^{m \times m} : W = W^T, diag(W) = 0\}, \tag{9.7}$$

in which W is a weight matrix uniquely describing a graph (and a graph Laplacian matrix L). The signal $Y \in \mathbb{R}^{m \times p}$ is a light field image, vectorized in such a way that each row represents one entire view, where $m = K \times N$ is the total number of views, and p the total number of pixels in one view. Increasing the parameter α enforces stronger connectivity in the graph, while decreasing β promotes sparsity.

In terms of our problem, ensuring a graph is connected is important, as it provides full flexibility in view selection. Indeed, if there were several separate connected components in the graph, view selection would need to provide samples from each of these components to ensure the reconstruction of the entire light field. While clearly a surpassable drawback, this would force the view selection to be dependent on the graph structure, complicating the procedure and making the problem no longer easily distributed. On the other hand, graph sparsity also represents an important parameter, as it reduces the overhead of transmitting the graph weights. As shown by Kalofolias et al. [Kalofolias, 2016], the problem in 9.5 is convex and has an efficient solution. The code is publicly available in the GSP toolbox [Perraudin et al., 2014].

Once the graph is encoded, an appropriate subset of views is selected. It is worth emphasizing that the graph learning step is carried out independently from the selection of the views and its compression, which brings several advantages. The first advantage is that different encoding solutions can be selected to efficiently compress the subset of views. Moreover, as the graph is always encoded losslessly and thus represents a fixed overhead, it can easily be included in any rate allocation problem. Another advantage is that several strategies can be implemented for the selection of the views to be encoded, depending on the use case. For a fixed bitrate, spatial resolution can be favored over angular resolution by selecting a smaller subset of views,

which will be compressed with a better quality. Conversely, sending a larger set of views will ensure a better angular resolution, while decreasing the overall quality of all the views. For instance, a progressive stream which would offer an increasingly superior angular resolution is straightforward to implement. Lastly, the learned graph structure can be used to select the subset of views in order to maximize the overall quality of the reconstructed light field, or to provide a trade-off between angular and spatial resolutions, depending on the desired application. As the dashed line in Figure 9.2 implies, one possibility is to exploit the knowledge of the estimated graph structure to select the views to be encoded, giving priority to more influential views to ensure a more faithful reconstruction.

The computational cost of learning the graph is $\mathcal{O}(m^2 p)$ for computing the distance between all views, and $\mathcal{O}(m^2)$ per iteration of the optimization problem. Taking into account the fact that the number of iterations i is limited [Kalofolias, 2016] and considering that in the majority of cases $p \gg i$, the overall complexity of learning the graph can be written as $\mathcal{O}(m^2 p)$. The cost of encoding the subset of views depends on the compression method of choice. Thus, it might be dominant in the overall compression scheme. For instance, the cost of learning the graph would be negligible with respect to the cost introduced by state of the art video codecs commonly used to encode the perspective images.

9.2.3 Decoder

After recovering the lossless graph and a lossy subset of views, the decoder exploits the graph to estimate the full light field. In order to recover the missing views, the decoder solves an optimization problem which enforces smoothness on the representative graph among the views. Namely, for a view selection matrix M, we want to solve:

$$\textbf{argmin}_X \quad tr(X^T L X) \tag{9.8}$$

$$s.t. \quad \hat{Y} = MX, \tag{9.9}$$

where $\hat{Y} \in \mathbb{R}^{m \times p}$ is a matrix containing decoded views in rows corresponding to one of the selected views, and zeros everywhere else. The view selection matrix M projects X to the space of selected view only, keeping only the values in corresponding rows. Specifically, it is an identity matrix with zeros on the diagonal for all indices corresponding to not selected views.

This problem can equivalently be written as:

$$\textbf{argmin}_X \quad tr(X^T L X) + \gamma \|\hat{Y} - MX\|_F^2 \tag{9.10}$$

with a tunable parameter γ that, if very small, allows changes also among the received views. It is worth noting here that received views went through a lossy compression. Therefore, it can be beneficial to allow small changes promoting smoothness on the graph, especially when the selected views are compressed with low bitrates.

Given the parameter γ, it is not difficult to see that the solution to problem 9.10 is given in closed form with:

$$\hat{X} = (M + \gamma L)^{-1} \hat{Y}, \tag{9.11}$$

which concludes the work of the decoder and gives the final estimation for the original light field image.

The view reconstruction step has a closed form solution, with the computational cost of $\mathcal{O}(m^3)$ for matrix inversion and $\mathcal{O}(m^2 p)$ for multiplication. As $p \gg m$ in most cases, the overall complexity can be written as $\mathcal{O}(m^2 p)$. However, depending on the choice of compression method for the subset of views, the cost of decoding the subset of views might be dominant in the overall compression scheme.

9.3 Validating experiment

In this section we give an overview of the validating experiment to test the performance of our solution. Specifically, we present the coding conditions and outline the codec configuration. We then introduce a brief description of the anchors and, lastly, delineate how the objective quality metrics are computed.

9.3.1 Coding conditions

In order to facilitate the comparison between the proposed approach and the state of the art in light field coding, the same coding conditions as defined in the ICIP 2017 Grand Challenge were adopted for this experiment [Viola and Ebrahimi, 2018a]. In particular, the following four light field contents were selected from the proposed lenslet dataset [Řeřábek and Ebrahimi, 2016]: *Bikes*, *Danger_de_Mort*, *Stone_Pillars_Outside* and *Fountain_&_Vincent_2* (see Figure 9.3).

The Light Field toolbox v0.4 was employed to obtain the 4D light field structure of perspective views [Dansereau et al., 2013, 2015]. Prior to the transformation, each 10-bit lenslet image was devignetted and demosaicked. A total of 15×15 perspective views were obtained from the lenslet image, each with a resolution of 625×434 pixels; however, only the central 13×13 views were selected to be encoded and evaluated, following the JPEG Pleno Common Test Conditions [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018a]. Color and gamma correction was applied to each perspective view prior to the encoding.

The same compression ratios defined for the Grand Challenge were selected for the evaluation of the proposals, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, $R4 = 0.005$ bpp. However, conforming to the JPEG Pleno Common Test Conditions [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018a], the bpp were computed as the ratio between the total number of bits used to

(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside (d) Fountain_&_Vincent_2

Figure 9.3 – Central perspective view from each content used in the validating experiment.

encode the content, and the total number of pixels in the entire light field, which in our case corresponds to $13 \times 13 \times 434 \times 625$ pixels.

9.3.2 Codec configuration

The graph was computed on the luminance values of the 4D light field structure. To reduce the overhead, only the luminance graph was transmitted, and it was used for the reconstruction of all YUV channels. Parameters $\alpha = 10^5$ and $\beta = 10$ were empirically chosen for the encoder, whereas for the decoder the parameter γ was set to 10^{-8} and $3 \cdot 10^{-4}$ for the luminance and for the chrominance channels, respectively. The weight matrix of the graph is symmetric and highly diagonally sparse. Therefore, the upper triangle of our weight matrix was rearranged using MATLAB function *spdiags* and losslessly compressed as a *mat* file. Information about the size of each graph can be found in Table 9.1.

For the experiment, the views composing the 4D light field structure were divided in two sets A and B, forming the views that would be compressed and transmitted alongside the graph, and the views that would be entirely reconstructed on the decoder side, respectively. A total of 85 out of 169 views were assigned to set A, whereas 84 views composed set B, as shown in Figure 9.4. The views in set A were subsequently converted to YUV color space, downsampled from 444 to 420, 10-bit depth, and compressed using the HEVC/H.265 reference software HM [ITU-T Q.6/SG 16 and ISO/IEC JTC 1/SC 29/WG 11], using profile Main10 and low delay configuration. The QPs were chosen to closely match the targeted compression ratio. A summary of the chosen QP and relative file size can be found in Table 9.1.

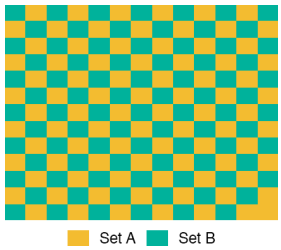


Figure 9.4 – Composition of set A and set B.

Table 9.1 – Size of the compressed bitstreams, and relative QPs for every content and compression ratio.

Content	Graph size	QP	Set A size	Total size	Compression ratio	Target size
Bikes	2489 B	12	3956.07 kB	3958.50 kB	0.707 bpp	4196.89 kB
		23	500.05kB	502.48 kB	0.090 bpp	559.59 kB
		31	110.28 kB	112.71 kB	0.020 bpp	111.92 kB
		41	23.72 kB	26.15 kB	0.005 bpp	27.98 kB
Danger_de_Mort	2646 B	14	4024.55 kB	4027.14 kB	0.720 bpp	4196.89 kB
		25	518.73 kB	521.32 kB	0.093 bpp	559.59 kB
		33	114.30 kB	116.89 kB	0.021 bpp	111.92 kB
		42	24.78 kB	27.36 kB	0.005 bpp	27.98 kB
Stone_Pillars_Outside	3306 B	12	3965.9 kB	3969.2 kB	0.709 bpp	4196.89 kB
		22	504.52 kB	507.74 kB	0.091 bpp	559.59 kB
		28	103.25 kB	106.47 kB	0.019 bpp	111.92 kB
		35	24.25 kB	27.48 kB	0.005 bpp	27.98 kB
Fountain_&_Vincent_2	1884 B	12	4225.2 kB	4227 kB	0.755 bpp	4196.89 kB
		24	493.56 kB	495.4 kB	0.089 bpp	559.59 kB
		31	114.03 kB	115.87 kB	0.021 bpp	111.92 kB
		40	26.25 kB	28.09 kB	0.005 bpp	27.98 kB

9.3.3 Anchor selection

The results of our coding approach were compared to the results obtained from HEVC/H.265 anchor used in the ICIP 2017 Grand Challenge [Viola and Ebrahimi, 2018a]. In the HEVC/H.265 anchor, the software implementation x265[1] is used to encode the perspective views, which were previously arranged in a serpentine order.

In addition, our results were compared to the best performing algorithm of the ICIP 2017 Grand Challenge[2], which defines a linear dependency among different views in the angular domain, called Linear Approximation Prior (LAP) [Zhao and Chen, 2017]. In their work, a

[1] https://www.videolan.org/developers/x265.html
[2] http://2017.ieeeicip.org/ChallengeAward.asp

subset of views is encoded using x265 and transmitted to the encoder along with the quantized linear coefficients. The rest of the views is then estimated using the LAP assumption.

Finally, the JPEG Pleno VM was used as third anchor [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018b, Astola and Tabus, 2018b]. The provided configuration for the four contents is used for the comparison. However, it should be noted that the configuration files are optimized for random access, which could negatively affect the performance of the codec in terms of objective quality.

9.3.4 Objective quality evaluation

To evaluate the performance of the proposed coding algorithm with respect to the anchors, PSNR and SSIM were selected from the literature as objective quality metrics, following the JPEG Pleno Common Test Conditions [ISO/IEC JTC 1/SC29/WG1 JPEG, 2018a]. In particular, every perspective view at indices (k,l) was converted to YUV color space, 10-bit depth, using the conversion matrix defined in Recommendation ITU-R BT.709.6 [ITU-R BT.709-6, 2015]. The metrics were then applied separately to the luma channel Y and for each viewpoint image, as follows:

$$PSNR_Y(k,n) = 10\log_{10}\frac{(2^{10}-1)^2}{MSE(k,n)}, \tag{9.12}$$

$$SSIM_Y(k,n) = \frac{(2\mu_I\mu_R + c_1)(2\sigma_{IR} + c_2)}{(\mu_I^2 + \mu_R^2 + c_1)(\sigma_I^2 + \sigma_R^2 + c_2)}, \tag{9.13}$$

in which $MSE(k,n)$ is the mean square error between the reference and the reconstructed view at indices (k,n), μ_I and μ_R are the mean values, σ_I^2 and σ_R^2 are the variances, and σ_{IR} is the covariance of the two perspective views in channel Y. PSNR was computed for channels U,V according to Equation 9.12, and a weighted average [Ohm et al., 2012] was obtained as follows:

$$PSNR_{YUV}(k,n) = \frac{6PSNR_Y(k,n) + PSNR_U(k,n) + PSNR_V(k,n)}{8}. \tag{9.14}$$

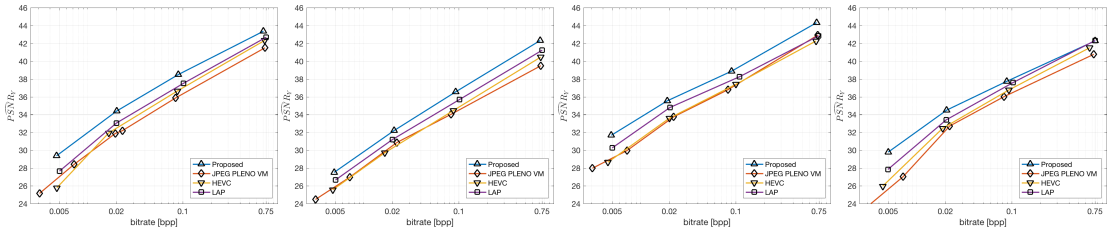
The average PSNR value for Y channel was then computed across the viewpoint images:

$$\widehat{PSNR}_Y = \frac{1}{(K-2)(N-2)}\sum_{k=2}^{K-1}\sum_{n=2}^{N-1} PSNR_Y(k,n). \tag{9.15}$$

Similarly, the average $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ values were computed. Additionally, Bjontegaard rate savings percentages and PSNR gains [Bjontegaard, Gisle, 2001] were computed with respect to all the anchors for $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$ values.

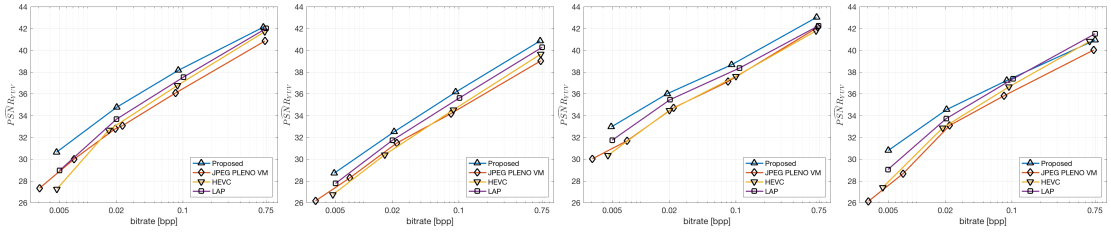
9.4 Results and discussion

Figure 9.5 shows the values of $\widehat{PSNR}_Y$ against the bitrate, separately for each content under examination. It can be seen how our proposal outperforms the anchors pretty consistently across different bitrates for contents *Bikes*, *Danger_de_Mort* and *Stone_Pillars_Outside*. For content *Fountain_&_Vincent_2* a notable gain can be observed for lower bitrates, whereas for high bitrates the performance is equivalent to codec LAP.



(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside (d) Fountain_&_Vincent_2

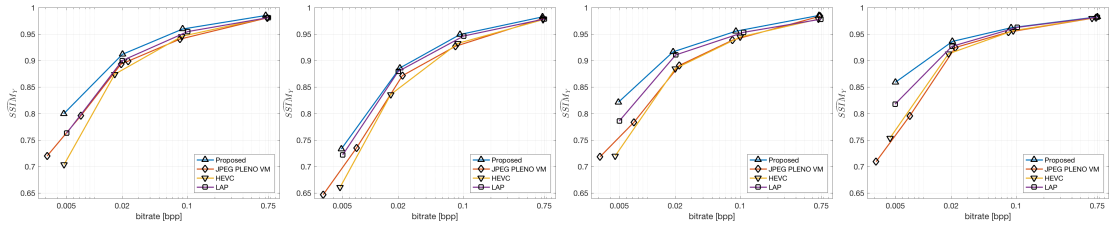
Figure 9.5 – $\widehat{PSNR}_Y$ vs bitrate for every content. The bitrate is shown in logarithmic scale to improve readability.



(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside (d) Fountain_&_Vincent_2

Figure 9.6 – $\widehat{PSNR}_{YUV}$ vs bitrate for different contents. The bitrate is shown in logarithmic scale to improve readability.

A similar trend can be observed for values of $\widehat{PSNR}_{YUV}$ (Figure 9.6). In particular, it is worth noting that, although the performance of our proposal remains consistently better than or equivalent to the anchors, a smaller gain in dB can be observed. This may be due to the fact that anchors HEVC/H.265 and LAP apply a chroma subsampling factor of 422, which leads to



(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside (d) Fountain_&_Vincent_2

Figure 9.7 – $\widehat{SSIM}_Y$ vs bitrate for different contents. The bitrate is shown in logarithmic scale to improve readability.

Table 9.2 – Bjontegaard rate savings with respect to the three anchors HEVC, JPEG Pleno VM and LAP, for all four contents, and on average.

	HEVC/H.265		JPEG PLENO VM		LAP	
	$\widehat{PSNR}_Y$	$\widehat{PSNR}_{YUV}$	$\widehat{PSNR}_Y$	$\widehat{PSNR}_{YUV}$	$\widehat{PSNR}_Y$	$\widehat{PSNR}_{YUV}$
Bikes	-46.22%	-43.10%	**-57.65%**	**-55.31%**	**-36.52%**	**-31.89%**
Danger_de_Mort	-47.53%	-45.37%	-50.97%	-47.64%	-30.14%	-26.14%
Stone_Pillars_Outside	**-53.46%**	**-50.02%**	-52.28%	-48.93%	-35.19%	-29.68%
Fountain_&_Vincent_2	-39.74%	-33.45%	-51.24%	-49.04%	-23.06%	-15.36%
Average	-46.74%	-42.99%	-53.04%	-50.23%	-31.23%	-25.77%

Table 9.3 – Bjontegaard PSNR difference with respect to the three anchors HEVC, JPEG Pleno VM and LAP, for all four contents, and on average.

	HEVC/H.265		JPEG Pleno VM		LAP	
	$\widehat{PSNR}_Y$	$\widehat{PSNR}_{YUV}$	$\widehat{PSNR}_Y$	$\widehat{PSNR}_{YUV}$	$\widehat{PSNR}_Y$	$\widehat{PSNR}_{YUV}$
Bikes	1.93 dB	**1.53 dB**	**2.41 dB**	**1.86 dB**	**1.33 dB**	**0.98 dB**
Danger_de_Mort	1.89 dB	1.51 dB	2.08 dB	1.56 dB	1.06 dB	0.76 dB
Stone_Pillars_Outside	**1.96 dB**	1.47 dB	1.97 dB	1.44 dB	1.13 dB	0.75 dB
Fountain_&_Vincent_2	1.40 dB	1.01 dB	1.94 dB	1.51 dB	0.72 dB	0.39 dB
Average	1.80 dB	1.38 dB	2.10 dB	1.59 dB	1.06 dB	0.72 dB

improved color fidelity. Moreover, the choice of using the luminance graph to reconstruct the chroma values may lead to a loss in performance in the proposed codec.

Values of $\widehat{SSIM}_Y$ why show that our proposal has similar performance with respect to the anchors for high bitrates (Figure 9.7). However, a significant gain can be observed for low bitrates with respect to the other codecs.

Bjontegaard rate savings results (Tables 9.2 and 9.3) show that our proposal achieves on average a 46.74% rate reduction and a PSNR gain of 1.80 dB for $\widehat{PSNR}_Y$ (42.99% and 1.38 dB for $\widehat{PSNR}_{YUV}$, respectively) when compared to HEVC/H.265. The maximum rate reduction is achieved for content *Stone_Pillars_Outside* (53.46% and 50.02% for $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$, respectively), while the minimum gain is achieved for content *Fountain_&_Vincent_2* (39.74% and 33.45% for $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$, respectively). Slightly higher rate reductions can be achieved in comparison to the JPEG PLENO VM (53.04% and 50.23% for $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$, respectively), for which bigger PSNR gains can also be observed (2.1 and 1.59 dB for $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$, respectively). When analysing the difference in performance between the JPEG Pleno VM and our solution, it should be noted that configuring the codec for random access may result in a sub-par performance, objective quality metric-wise. Moreover, more recent versions of the verification software, which implement different coding strategies such as 4D-DCT, might achieve a better performance on the dataset. Smaller, but still significant gains can be seen by using our solution with respect to LAP codec, with a rate reduction of 31.23% and 25.77% on average for $\widehat{PSNR}_Y$ and $\widehat{PSNR}_{YUV}$, respectively, and a PSNR gain of 1.06 and 0.72 dB.

Results show that light field compression efficiency can benefit from sending only a subset of the perspective views and reconstructing the entire 4D light field at the receiver side, as shown by the superior performance of both the proposed solution and the LAP codec with respect to the HEVC/H.265 anchor. In particular, both LAP and our proposed solution rely on sparsely capturing the similarities among the perspective images to aid in the reconstruction process at the decoder side. However, in the LAP algorithm the reconstructed images are seen as a linear combination of only the views that have been encoded and sent, thus disregarding the correlation among the views that need to be reconstructed. On the other hand, our approach encodes all the dependencies in the 4D light field, regardless of the set they belong to. Moreover, whereas the coefficients of the linear dependency among views are quantized in the LAP scheme, the graph weights are losslessly compressed in our solution to improve reconstruction quality. Results show that this approach achieves a superior performance in reconstructing the 4D light field.

9.5 Conclusions

In this work we presented a new approach to compress light field images based on a graph learning technique. We demonstrate its theoretical soundness, as well as its application to image coding. Our validating experiment shows that sensible gains can be achieved by using our solution against state-of-the-art encoders.

While graph signal processing techniques have been used in other works to improve the coding efficiency for light field images, the construction of the graph is usually imposed on the data. In our work we focus on learning the graph in order to faithfully capture the similarities among perspective views. Although the graph learning technique used on this paper is lifted from

the literature, the way it is employed to estimate the dependencies among views has not been used in the past. Moreover, other work in literature considers each pixel as a vertex in the graph, resulting in the construction of very large graphs that weight heavily on the final bitrate. By using each view as a node, we considerably reduce the size of the graph while retaining its capability to capture variations among different views.

The main contributions of this chapter can be summarized as follows:

- We designed a lightweight predictive scheme to capture the similarities and redundancies in the light field data. The predictive scheme uses known graph learning techniques to infer the similarities among the views that compose the light field structure. Unlike previous work in graph signal processing for light field compression, each view is selected as a node in the graph, which allows to keep the overhead small while ensuring good predictive power. The graph is constructed so to promote smoothness among the views, while enforcing connectedness and sparsity.
- We constructed a compression algorithm that uses the aforementioned predictive scheme to recover the entire light field from an arbitrarily chosen subset of views. The compression algorithm generates the graph from the entire light field and compresses it in a lossless fashion. A subset of views is selected and subsequently encoded. At the decoder side, the graph and the encoded views are used to reconstruct the original light field data. As graph connectivity is enforced, the views to be encoded can be selected with full flexibility. Moreover, as the graph construction is independent from the selection and encoding of the subset of views, any state-of-the-art compression algorithm can be used in the encoding.
- We tested our solution on a widely used light field dataset and we compared our results with state-of-the-art light field compression algorithms, using well-know image quality metrics. We showed that sensible gains can be achieved by using our solution with respect to the other algorithms.

Possible extensions of this work include improving coding efficiency by implementing a more efficient selection of encoded views based on graph structure, and improving color performance by incorporating chroma information in the graph weights. Moreover, the selection of nodes for the construction of the graph can be modified to further capture the variation across different views, for example, by segmenting each view and assigning a node to each segment.

Part IV
Conclusions and future work

10 Conclusions

10.1 Outcomes and accomplishments

This dissertation presented the results of investigating several aspects of compression and visual quality assessment for light field contents. Our work was divided into three parts, corresponding to three main areas of interest:

1. **Methodologies and scenarios for visual quality assessment of light field contents.** The part focuses on the theoretical aspects of visual quality assessment, such as the validity of single-image, interactive and passive evaluation for light field contents, analysis of user behaviour, and cross-display differences. The concepts demonstrated in this part, while deeply rooted in quality assessment of compression efficiency, can easily be generalized for the assessment of light field contents under any type of distortions, and offer useful guidelines for future design of subjective quality methodologies.

2. **Analysis and comparison of compression solutions for light field contents through visual quality assessment.** This part focuses on the performance of several algorithms for light field compression, and how the adoption of both established video-encoding solutions and state-of-the-art compression architectures can impact the visual quality, based on different types of representation. The results presented in this part offer an impartial, reliable benchmarking of the state of the art in light field encoding, and demonstrate how the choice of representation models and rendering technology can sensibly affect the performance of the selected compression solution.

3. **Towards new compression solutions for light field contents.** The last part is directed at exploiting graph learning techniques to lift useful interdependency information from the light field data, which can be used to reduce the amount of information that needs to be encoded and transmitted. Results show that view estimation is a viable and promising solution to reduce the volume of data without compromising the visual quality.

In the following sections, we outline the contributions that have been presented in the three

main areas of interest.

10.1.1 Methodologies and scenarios for visual quality assessment of light field contents

- We consider the merits of single-image assessment in estimating the visual quality of light field contents through image-based rendering. In particular, we employ perspective and refocused views in our subjective evaluations, and we inspect the similarities in score distributions both within and across rendering groups. Our results suggest that selecting more than one view in each rendering group does not lead to significantly different results. Thus, we conclude that the additional strain, imposed by increasing the number of stimuli under test, is not justified by the corresponding scores. We also note that the differences in score distribution between perspective and refocused views are statistically relevant; in particular, refocused views generally receive harsher scores than their perspective counterpart, and fail to reach transparent quality, even when no compression is applied.
  The main takeaway from our study would be to employ only one view per rendering group, considering that multiple views belonging to the same rendering group received similarly distributed scores in our test. However, the fact that the scores assigned to different rendering groups were statistically different should be carefully considered when deciding how to perform the test. As the visual quality of refocused views struggled to reach transparent quality even in the absence of compression artifacts, we are prone to suggest perspective views should be preferred when designing a subjective test. However, given that the visual quality of each perspective view may considerably vary, depending on which compression algorithm is applied, we are reluctant to recommend the use of single-image assessment as a one-size-fits-all solution for subjective evaluation. Combining several perspective views in one single stimulus to be assessed, whether it would be with interactive or passive tests, could be beneficial for the evaluation of compression solutions for which the quality of perspective views is not homogeneous.

- We perform a comparison between subjective methodologies for light field quality assessment. The first methodology allows the users to interact with the content by changing rendering parameters, while the other favors a passive animation which ensures the same experience will be given to all the users. Results demonstrate that, while the two methodologies are strongly correlated, they are not statistically equivalent. Moreover, smaller CIs are associated with the scores collected through the passive methodology, despite the fact that a semi-controlled crowdsourcing platform (which normally leads to higher variance) is used. Thus, the passive approach has more discriminative power when compared to the interactive one.
  Our main recommendation would be to adopt the passive methodology in subjective tests when it is crucial to be able to distingush among different tested solutions, since providing a uniform visualization experience through the use of a pre-recorded

animation leads to smaller CIs and more differentiation among the scores. However, interactive assessment should not be brushed off, as it offers a more realistic scenario for the consumption of light field contents. More analysis is needed to assess how strengths of both approaches can be combined to offer an immersive scenario for experiencing light field contents, while providing a consistent experience for all users.

- We analyze how user behavior is affected by different compression distorsions, by performing an interactive subjective test where users' actions are recorded. We define new procedures to aggregate the user behavior data into time of interaction, and we perform correlation between the subjective scores and the time of interaction to extract patterns and trends. Results show that the total time of interaction is strongly correlated with the subjective scores; in particular, subjects are prone to spending more time on good-quality contents, as opposed to very distorted ones, regardless of the presentation order.
  Our work lays the basis for implicit assessment of light field contents, using time of interaction as indicator for visual quality. However, further tests are needed to confirm whether time of interaction remains a good predictor for quality scores when different tasks, or free viewing, are employed. To help foster research in the field, we provide the open-souce framework in Annex B, and the data we used and collected for our research, including the compressed contents, the MOS scores and the user behavior information, in Annex A.

- We test different scenarios for visual quality assessment of light field contents, rendered using multi-layer technology. We define two variants for the DSIS methodology, and we test both variants on different laboratory settings and using both a prototype display and a simulator for 2D screens. In particular, we first compare two sets of scores obtained using the same methodology on the simulator software, in different laboratory settings, to account for cross-cultural difference. Our findings show that the two are highly correlated, although the scores obtained in one laboratory setting are consistently higher with respect to the other. Secondly, we perform a comparison between the two DSIS variants we defined, proving that one of the variants leads to less uncertainty for high MOS scores with respect to the other. Finally, we examine the correlation between the scores collected using the prototype display and the simulator, respectively. Results demonstrate that statistically different, uncorrelated results are obtained with the two visualization setups. Notably, the score distribution among the two sets shows that, for the case of the multi-layer display, the method employed for generating the images to be used for the rendering has a bigger impact on the final MOS scores.
  Several outcomes can be sorted from our test. The first and most obvious, is that cultural biases should be taken into account when evaluating the visual quality of light field contents. Secondly, the performance of side-by-side evaluations in discerning among competing solutions can be improved, especially at near-transparent levels of quality, by considering other variants that exploit the perception of temporally-variant features, like our proposed variant for the DSIS methodology. Lastly, the impact of hardware

limitations and diminished QoE on the distribution of the collected scores should be taken into consideration when designing a subjective evaluation campaign. The poor correlation between the scores obtained with the simulator and the prototype display show that an ideal rendering setting, where no distortions are added to the final rendered content, may misrepresent the way contents are perceived in noisy scenarios.

10.1.2 Analysis and comparison of compression solutions for light field contents through visual quality assessment

- We provide the results of the objective and subjective quaity assessments conducted during two preeminent grand challenges on light field compression. The outcomes of the evaluation campaign are thoroughly analyzed, so that the best-performing approaches can be identified for future design of compression solutions.
  Although a preliminary performance evaluation is usually given along with any new compression solution, the coding conditions are seldom uniform among different works. Thus, judging the efficiency of any given compression algorithm with respect to the state of the art becomes a delicate task, as it becomes near-impossible to perform a benchmarking in a variety of conditions that would comprehensively test the performance under various levels of stress. Our main contribution, in this case, is providing an exhaustive examination of state-of-the-art solutions in light field coding using an expert-selected range of coding conditions, to facilitate the assessment of new proposals for efficient light field coding. To further help research in light field compression, we provide a dataset containing the objective and subjective results relative to the anchor data, along with selected submissions to the latest grand challenge. More information can be found in Annex A.

- We define and examine two coding approaches for lenslet-based light field compression: one which encodes the raw data after minimal pre-processing, and the other operating on the 4D light field structure level. Moreover, we investigate the impact of applying chroma subsampling on the raw data. We select five algorithms for light field compression that comply to either of the approaches, and we carry out a comparison through both objective and subjective means. Our results show that a superior performance is achieved when the 4D light field structure, which allows to exploit the redundancies among neighboring views, is encoded. Furthermore, we demonstrate that one post-processing procedure, which is applied during the transformation to 4D light field structure, is able to propagate errors from the raw data, such as chroma subsampling, leading to undesired results and affecting the quality of the final rendered product.
  One of the main accomplishments of our work is to determine how commonly used rate-reduction methods, such as chroma subsampling, cannot be applied on the raw lenslet data, if we want to preserve the visual quality of the final result. Moreover, if rate-optimization is carried out on the raw data level, as commonly done in intra lenslet compression solution, unwelcome effects, such as a sharp loss in performance, could

be observed on the post-processed 4D light field. Thus, if an intra approach is pursued, we recommend incorporating information about the post-processing procedures in the coding architecture, so that the process is optimized according to the final rendered quality.

- We outline three compression strategies for light field contents that will be rendered using multi-layer displays. While one solution is renderer-agnostic and simply performs the compression on the 4D light field, the other two take advantage of two methods to generate the layer patterns that will be used in the multi-layer display for the rendering, thus reducing the amount of data that needs to be encoded. We perform objective and subjective quality assessment; for the latter, both a prototype multi-layer display and a simulator software are used. Results show that, while one approach is superior to all others in all the scenarios, contradicting outcomes can be found when the simulator or the multi-layer display are employed for the test. In particular, scores obtained with the simulator offer more differentiation among different compression ratios, with respect to both objective quality metrics and scores from the test conducted using the multi-layer display.
  Our work demonstrates how different methods for the generation of layer patterns for multi-layer rendering have a strong impact on the final rendered quality of both compressed and uncompressed contents. Considering how transparent levels of quality are never achieved for one of the methods, we recommend incorporating subjective and objective quality assessment information to optimize the generation of layer patterns, to improve the final rendered quality. Moreover, when multi-layer rendering is considered, we disfavor the use of objective quality metrics for predicting the visual quality of light field contents under compression artifacts, as they were shown to discriminate more on the basis of the method for generating the layer patterns. The same applies for prototype display, whose hardware limitations strongly affect the subjective score distributions.

10.1.3 Towards new compression solutions for light field contents

- We propose a new graph-based compression algorithm for light field contents. Our algorithm uses graph learning to collect information about the dependencies among views; the information is losslessly transmitted, and is used to reconstruct the 4D light field from a subset of encoded views. In order to evaluate the compression efficiency of our proposed solution, we adopt common test conditions from the JPEG standardization body, and we compare our results with state-of-the-art algorithms. Results show that notable gains can be achieved by adopting our encoding scheme.
  As the main outcome of this work, we propose an efficient method to generate lightweight disparity data from the light field structure. Sparsity and connectivity are enforced when generating the disparity information, and the resulting graph is efficiently stored exploiting the diagonal structure of its weights. To promote further inquiries on the topic, the code used to generate the graph is made publicly available.

10.2 Limitations and future prospects

The work presented in this dissertations could be extended and improved in several aspects. We summarize here how we envision it to be continued:

- In Chapter 3, our analysis of single-image assessment was conducted on only one methodology, namely DSCQS. It could be interesting to measure whether the same effects can be found when other single- or double-stimulus methods are employed, and in presence of alternative compression solutions and coding condition.
- In multiple instances, we have adopted interactive assessment for evaluating the visual quality of light field contents (see for example Chapter 3, 4 and 7). However, it should be noted that the parameters for the rendering were chosen and applied offline, as the delay induced by the real-time rendering operations would severely compromise the user experience. Moreover, the current interactive rendering setup is suitable for light field images whose angular and spatial resolution falls within a modest range, such as contents acquired with plenoptic cameras. Higher spatial and angular resolution could potentially not be supported by the rendering system, because of delay and memory constraints. In this case, lossless compression schemes to reduce the strain on memory consuption should be designed and employed. Furthermore, ad-hoc rendering systems could be envisioned to sustain the interactive methodology for any type of light field content.
- Our evaluation of compression solutions and subjective quality assessment methodologies for multi-layer rendering, which we conducted in Chapters 5 and 8, can be extended in several directions. First, a multi-layer setup with a different number of layers could be considered. Secondly, the work can be broadened by considering video contents, instead of still images, as they are supported by both the prototype multi-layer display and by the simulator software. Furthermore, a comparison between the QoE associated with different visualization approaches, as image-based rendering, MR-DIBR and multi-layer rendering, could be conceived to assess which rendering technology leads to the best results.
- As specified in the previous section, in Chapter 4 we lay the basis for implicit assessment of light field contents. The research can be extended by comparing it with the interaction patterns obtained with other types of task-dependent evaluations, in order to measure whether the same correlation with visual quality can be found. Moreover, different types of distortions besides compression artifacts could be considered, to evaluate whether the phenomenon is circumscribed to encoding alterations alone. Alternative methodologies can be tested to see if the results can be generalized for other types of comparison.
- Our work on the impact of different coding approaches and rate-reduction techniques on the visual quality of light field contents, which we conducted in Chapter 7, was

focused solely on lenslet-based light field contents, and specifically on images acquired with an unfocused plenoptic camera. Thus, the work can be extended by considering different acquisition technologies, in particular focused plenoptic cameras and coded aperture solutions. The testing conditions could be broadened by considering a wider range of compression ratios, not to mention alternative compression algorithms. Moreover, single-stimulus subjective quality assessment methodologies could be employed to measure the effect of chromatic distortions on the perceived quality in an absolute scale.

- Our new compression solution for encoding disparity information, presented in Chapter 9, was only tested on lenslet-based light field contents, and in fact works best when the angular sampling is quite dense. Future research should focus on extending the graph learning technique to be suitable for a sparser or irregular sampling grids, for example by constructing more than one graph, to capture both fast and slow transitions among perspective views. Moreover, the performance on chroma channels can be improved by incorporating color information in the graph learning process. Furthermore, in our work a smoothness prior was selected because of the characteristics of the data; more complex prior information could be included in the graph construction.

The main objectives of this dissertation were to provide an analysis of various methodologies for quality assessment of light field contents, to evaluate the compression capabilities of various encoding solutions, and to propose a new method to improve the coding efficiency for light field contents. Further research is needed to improve the quality of experience for the users, to achieve a widespread adoption of light field imaging for the visualization and consuption of 3D scenes in immersive scenarios.

Annexes

A A dataset for visual quality assessment of light field images

Disclaimer: This chapter was adapted from the following article, with permission from all co-authors and publishing entities:

Viola, Irene, and Touradj Ebrahimi."VALID: Visual quality Assessment for Light field Images Dataset" in 10th International Conference on Quality of Multimedia Experience (QoMEX). ©2018 IEEE.

Light field imaging offers new ways of interaction with real-life scenarios in an immersive environment. However, the large volume of data generated in the acquisition process represents a challenge in terms of storage and transmission. The design of new compression solutions relies on subjective and objective visual quality assessment to efficiently reduce the amount of data while preserving both perceptual and immersive features. However, subjective assessment is costly and time consuming. Thus, comprehensive datasets for visual assessment of light field contents under compression artifacts are indispensable.

Several light field image datasets have been proposed in the past, comprised of both synthetic and natural scenes [Wetzstein, 2010, Laboratory, 2004, Řeřábek and Ebrahimi, 2016], and for object recognition and saliency map estimation [Ghasemi et al., 2014, Li et al., 2014a]. However, none of the datasets includes objective and subjective quality scores for compression-like artifacts. Paudyal et al. [Paudyal et al., 2017b] propose a so-called SMART dataset including several light field images compressed at various bitrates, along with the annotated subjective scores. However, the proposed compression solutions only consider intra-based approaches to encode light field images, which were proven to be subpar with respect to pseudo-sequence based approaches [Viola et al., 2017a]. Moreover, the subjective methodology that is used to collect the scores presents light field contents as conventional 2D images, which admittedly disregards any problem that may arise in the encoding of the depth information. Additionally, no data about the participants is provided, and the results are already processed in BT scores with respective CIs, so it is not possible to perform outlier detection or use a subset of the rates.

(a) I01 (b) I02 (c) I03 (d) I04 (e) I05

Figure A.1 – Central perspective view of each content from the proposed VALID dataset. ©2018 IEEE

In this annex we present a new dataset for visual quality assessment of light field images (VALID). The dataset is composed of uncompressed and compressed contents on various bitrates using four compression solutions. Objective quality results based on PSNR and SSIM metrics are provided, along with subjective quality assessment scores obtained using three different methodologies. Two visualization arrangements with different color bit depth are used. A summary of the contents of the dataset can be found in Table A.1.

A.1 Dataset description

A.1.1 Content and bitrate selection

Five lenslet-based light field images were chosen from a publicly available light field image dataset, namely I01 = *Bikes*, I02 = *Danger_de_Mort*, I04 = *Stone_Pillars_Outside*, I09 = *Fountain_&_Vincent_2* and I10 = *Friends_1* [Řeřábek and Ebrahimi, 2016]. The images were carefully selected from those commonly used in literature [Viola et al., 2017a, Ahmad et al., 2017, Tabus et al., 2017], to provide a variety of scenarios, containing a wide range of details that would be challenging for the compression algorithms in terms of texture and disparity encoding. From each lenslet image, 15×15 perspective views of 625×434 pixels and depth of 10 bits per color channel were obtained, using the Light Field toolbox v0.4 [Dansereau et al., 2013, 2015]. The central perspective view from the contents is depicted in Figure A.1. In order to provide compression distorsions at different levels of visual quality, four bitrates were selected: 0.75 bpp, 0.1 bpp, 0.02 bpp, 0.005 bpp. The values are obtained by dividing the size of the compressed bitstream over the size of the uncompressed raw images (5368×7728 pixels).

A.1.2 Encoding solutions and data preparation

A total of five solutions were adopted to compress the light field contents. Two popular video encoders, HEVC and VP9, were selected to encode the perspective views from the light field contents as pseudo-temporal sequences. For HEVC, the software implementation x265[1] was

[1] https://www.videolan.org/developers/x265.html

Table A.1 – Summary of contents for the VALID dataset. ©2018 IEEE

Content	Bitrate (bpp)	Objective quality metrics	Bit depth	Display	Size	Resolution	N_P	N_R	Methodologies	Codecs
I01 I02 I04 I09 I10	0.75 0.1 0.02 0.005	$PSNR_Y$ $PSNR_{YUV}$ $SSIM_Y$ $SSIM_{YUV}$	8 bit	Samsung SyncMaster2443	24in	1920×1200	81	11	Passive Interactive Passive and interactive	HEVC VP9
			10 bit	Eizo ColorEdge CG318-4K	31.1in	4096×2160	97	-	Passive	HEVC VP9 [Zhao and Chen, 2017] [Ahmad et al., 2017] [Tabus et al., 2017]

used, with the Main10 profile. For VP9, the official implementation was employed[2]. The QP and the target bitrates were selected to match the desired compression ratios for HEVC and VP9, respectively. To be used for the encoding, the perspective views were padded with black pixels, converted to YCbCr format and downsampled from 444 to 422, 10-bit depth. They were then arranged in a pseudo-temporal arrangement following a serpentine order. Only the central 13×13 perspective views were encoded.

Additionally, three state-of-the-art algorithms were selected from the literature to provide up-to-date results on light field compression. In [Zhao and Chen, 2017] authors encode a subset of the perspective views using HEVC, adopting a linear approximation prior to estimate the non-encoded views. In [Ahmad et al., 2017] authors arrange the perspective views into a multiview structure that can be exploited by the corresponding extension of HEVC, namely MV-HEVC. They also propose a rate allocation scheme to progressively assign the QPs in order to optimize the performance. In [Tabus et al., 2017], a lenslet-based compression solution that uses depth, disparity and sparse prediction information to reconstruct the final set of views is designed. The scheme can be configured to improve the reconstruction by allocating a fraction of the bitrate to the encoding of the lenslet image using JPEG 2000, or to allow random access by encoding a subset of views.

A.1.3 Output bit depth

Two output bit depths were considered for the objective and subjective assessments. Initially, 10 bits per color channel (the original bit depth of the images) were used to test the encoding solutions. All codecs were considered for the assessments. Additionally, the output of the encoding algorithms was converted to 8 bits per color channel, to ensure compatibility with the majority of consumers' devices and rendering softwares. Multiple methodologies were assessed to give an overview of different visualization and interaction approaches. For the 8 bit depth case, only HEVC and VP9 were used.

A.1.4 Objective quality metrics

PSNR and SSIM were selected from the literature to provide objective assessments of the visual quality of the contents. The metrics were applied separately to each luminance and chrominance channels Y, U, V and to each perspective view (k, l), where $k = 1, ..., K$, $l = 1, ..., L$ and $K = L = 15$ represent the total number of perspective views, as generated from the toolbox. $PSNR_{YUV}$ and $SSIM_{YUV}$ were computed by means of a weighted average, assigning factor 6 to the luma channel, and factor 1 to each chrominance channel, as defined in [Ohm et al., 2012].

The mean across the viewpoint images was also computed to have the average PSNR values

[2] https://www.webmproject.org/vp9/

for Y channel:

$$\widehat{PSNR}_Y = \frac{1}{(K-2)(L-2)} \sum_{k=2}^{K-1} \sum_{l=2}^{L-1} PSNR_Y(k,l), \tag{A.1}$$

Similarly, $\widehat{SSIM}_Y$, $\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_{YUV}$ were computed.

For the sake of completeness, the objective quality metrics were calculated on both the 10-bit and 8-bit outputs.

A.1.5 Subjective methodologies and test conditions

The subjective quality evaluations were conducted in a laboratory for subjective quality assessment, which was set up according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012], and equipped with adjustable neon lamps of 6500 K color temperature. The color of the background walls was mid grey, and the illumination level measured on the screens was 15 lux. The distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content, conforming to requirements in ITU-R Recommendation BT.2022 [ITU-R BT.2022, 2012]. Subjects were allowed to move further or closer to the screen. Specification about the display size and resolution can be found in Table A.1. All monitors were calibrated according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level of 0.2 cd/m^2.

Different subjective methodologies were considered based on the output bit depth. For the 10-bit output depth, the encoding solutions were tested using a "passive" methodology, using $N_P = 97$ perspective views at a rate of 10 frames per second, as recommended in [Viola et al., 2017b]. However, no refocusing was applied on the views ($N_R = 0$), to exclusively compare the outcome of the encoding algorithms. The total length of the animation was 9.7 seconds. A comparison-based adjectival categorical judgement methodology with a 7-point grading scale was selected, according to ITU-R Recommendation BT.500-13 [ITU-R BT.500-13, 2012]. Each stimulus was displayed alongside the uncompressed reference in a side-by-side arrangement. Participants were asked to compare the quality of the test stimuli with respect to the uncompressed reference and rate it on a scale from -3 (much worse) to +3 (much better), 0 indicating no preference.

For the 8-bit output depth, three methodologies were adopted, to test the impact of different visualization and interaction approaches on the collected subjective scores. Namely, "interactive" and "passive" approaches were implemented to collect the scores, and they were subsequently combined ("passive and interactive" approach) to offer interaction while improving the consistency of the results, as suggested in [Viola et al., 2017b]. In particular, for the "passive and interactive" approach, the participants were shown an animation of the images

under test, and could not interact or score before the animation was concluded. To ensure a smooth interaction experience without unwanted distorsions, only the central 9×9 views were used for the tests ($N_P = 81$). Additionally, $N_R = 11$ refocused views were created following [Viola et al., 2017b]. A DSIS methodology with side-by-side visualization and 5-point grading scale, from 5 (imperceptible) to 1 (very annoying), was selected for all three methodologies. For the "passive" and "passive and interactive" methodologies, the perspective views were shown as an animation, at a rate of 10 frames per second, followed by the refocused views, going from foreground to background and from background to foreground at a rate of 4 frames per second, as suggested in [Viola et al., 2017b]. The total length of the animation was 13.6 seconds. The "interactive" and "passive and interactive" methodologies were implemented using the framework proposed in [Viola and Ebrahimi, 2017], to allow subjects to engage with the perspective and refocused views.

In all the experiments, the position of the reference was fixed for the duration of the test, and participants were informed of its position on the screen. A training session with four training samples was established before the experiment, composed of one additional content compressed at various bitrates. The order of the stimuli was randomized for each participant, and the same content was never shown twice in a row. All subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively. Information about the age and gender of the participants is provided separately for each test. For all the evaluations, subjective scores are provided for each stimulus and for each participant. Additionally, for the "interactive" and "passive and interactive" methodologies, the tracking values from the animation are additionally provided for each subject and for each stimulus, to help analyse user behavior.

The dataset can be found in: https://mmspg.epfl.ch/VALID. Terms and conditions are given with the dataset. In case of use of the contents of the dataset for any purpose, as well as when presenting and publishing results based on the dataset or any of its parts, a reference to [Viola and Ebrahimi, 2018b] should be provided.

B A new framework for interactive quality assessment with application to light field coding

Disclaimer: This chapter was adapted from the following article, with permission from all publishing entities:

Irene Viola, Touradj Ebrahimi, "A new framework for interactive quality assessment with application to light field coding," Proc. SPIE 10396, Applications of Digital Image Processing XL, 10397F (19 September 2017). DOI: https://doi.org/10.1117/12.2275136

A wide range of possibilities are present in image-based rendering of light field contents. For example, it is possible to combine different perspective views to change the focal plane, in an interactive way. These peculiarities have to be mirrored in the methodology used for subjective quality evaluation.

Assessing the way users engage with light field contents plays a major role on how those methodologies are designed and used. However, user behaviour when engaging with light field content has not yet been studied in details.

In this chapter, we propose a new framework for visual quality assessment of light field contents, which allows for interaction with the content and assessment of user experience by tracking user behaviour information. Such information can be subsequently used to further analyze patterns in user interaction. Applications of user interaction information can be found in development of new objective quality metrics, new subjective methodologies and new perceptual coding algorithms.

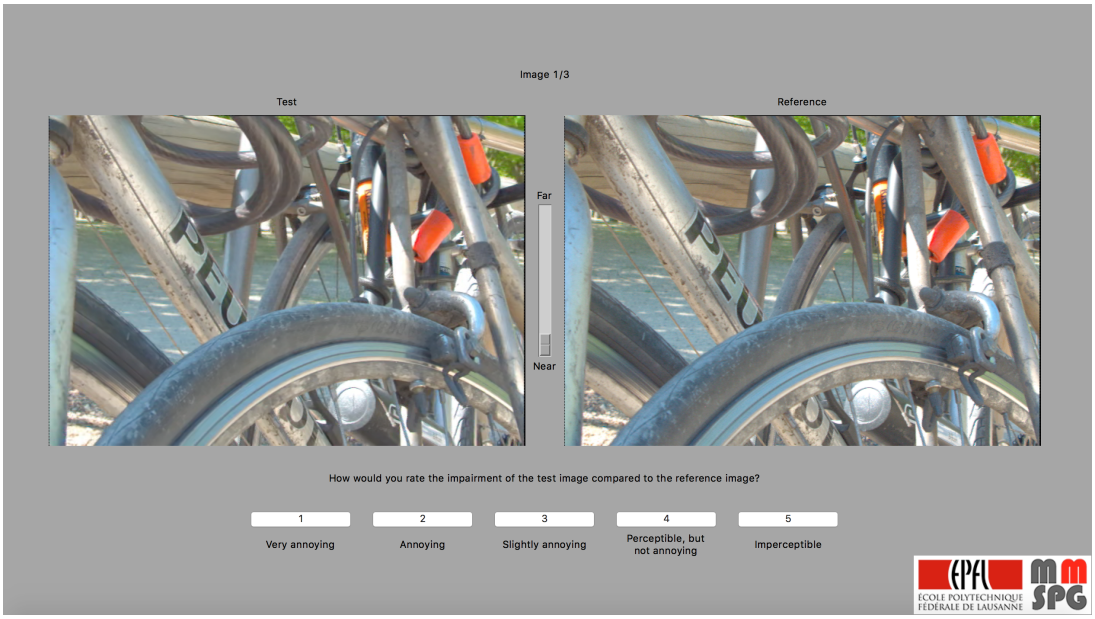


Figure B.1 – Example of evaluation interface screen.

B.1 Proposed framework

The framework proposed in this paper provides a tool to further understand the impact of user behaviour in quality assessment of image-based light field rendering. It consists of a software application for quality assessment of light field contents that enables interaction with the content while keeping track of what the user chooses to visualize. In its current version, the stimuli-comparison method DSIS is implemented, although the implementation of other methodologies through the framework is straightforward. A graphical interface allows interaction with light field contents in a real-time scenario, by enabling the change of point of view (pespective) and the choice of different focal points (refocus) from a predefined set. Figure B.1 shows an example from the framework.

The software takes as input two collections of perspective views in *png* file format, one serving as test and the other serving as reference. The perspective views are then assembled to form the light field content, composed of $U \times V$ images of resolution $W \times H$. Additionally, the software can receive as input a set of S images rendered from the light field content at different focal points, which we will refer to as refocused views, and a depth map D that can be used to access the refocused views. Both refocused views and depth map are saved in *png* file format. Test and reference materials must have the same resolution; moreover, they need to be rendered with the same parameters.

The central perspective view from the light field content taken as input is displayed as default

for both reference and test materials. By click-and-drag inside the rendered images, the user can change the perspective view, which is rendered in real time. A slider between the two rendered images allows access to the refocused views. Labels on each side of the slider indicate if the content will be refocused on the foreground or on the background. Additionally, the refocused views can be accessed by double clicking on any point of the image. In this case, the depth map is used to map the refocused views to each region in the scene. By clicking and dragging in any point of the rendered image, the user can return to visualize the perspective views. The two contents are rendered simultaneously and they are perfectly synchronized, so the displayed views are rendered with the same parameters. A panel on the bottom of the screen shows the possible scores for the test material. As soon as the user selects one option, the screen updates with the new test material.

The results of the evaluation are saved in one text file. Another text file provided as output records every perspective and refocused view that was accessed by the user, in access order. The start and end times of visualization of each view are recorded, along with the total display time.

A *python* implementation of the proposed framework can be found at the following link: https://github.com/mmspg/light-field-tracking. It is free to use, modify or redistribute, according to the GNU license. In case of use of the software for any purpose, publishing or use of any updates and variations based on it, as well as when presenting and publishing results based on the software, a reference to [Viola and Ebrahimi, 2017] should be provided.

C A comprehensive framework for visual quality assessment of multi-layer light field displays

Disclaimer: This chapter was adapted from the following article, with permission from all publishing entities:

Irene Viola, Keita Takahashi, Toshiaki Fujii and Touradj Ebrahimi, "A comprehensive framework for visual quality assessment of light field tensor displays," in Electronic Imaging 2019, Society for Imaging Science and Technology (IS&T), 2019.

In this chapter, we present a framework to conduct quality assessment of light field contents rendered through a tensor display simulator using 2D screens. Through a GUI, the layer patterns composing the multi-layer tensor displays are simulated in a 3D environment. By interacting with the mouse, users can experience the light field from different points of views.

C.1 Proposed framework

The framework proposed in this paper provides a tool to assess the quality of light field contents rendered through the use of multi-layer tensor displays. It consists of a software application for quality assessment of light field contents that enables visualization from different points of views, while keeping track of both the given ratings and the total time of interaction.

A graphical interface based on the software proposed in [Takahashi, 2018], simulates the multi-layer structure of light field tensor displays. The layer patterns are given as input to be directly visualized using the interface, along with a file specifying the parameters for the rendering, such as the horizontal and vertical angular resolution (i.e., the number of perspective views) of the input light field, and the number of layers composing the simulated display. The layer patterns are always displayed in their original resolution. By clicking and dragging, users can physically alter the visualization angle of the simulated display on the screen, thus accessing

Figure C.1 – Example rendering of the input stimuli with the proposed GUI, using double stimulus methodology with side-by-side display.

different points of views. The viewing angles are limited by the number of layers and the angular resolutions of the input light field, to ensure only properly rendered points of view are accessible. In particular, denoting V_x and V_y as the number of perspective views of the input light field in the horizontal and vertical dimension, respectively, and L as the number of layers in the simulated display, the maximum viewing angle θ_x and θ_y in the horizontal and vertical direction, respectively, can be defined as such:

$$\theta_x = \arctan \frac{\alpha \left\lfloor \frac{V_x}{2} \right\rfloor}{L}, \tag{C.1}$$

$$\theta_y = \arctan \frac{\alpha \left\lfloor \frac{V_y}{2} \right\rfloor}{L}. \tag{C.2}$$

Parameter α depends entirely on the specifications of the 2D monitor used to display the simulation:

$$\alpha = \frac{\sqrt{W^2 + H^2}}{\sqrt{w^2 + h^2}}, \tag{C.3}$$

in which W and H represent the screen size in meters, while w and h represent the screen resolution in pixels.

The graphical interface has been adapted to be used for subjective quality assessment. In accordance with the ITU-R recommendations [ITU-R BT.500-13, 2012], both single and double stimulus methodologies can be used for the subjective evaluation. For the double stimulus methodology, both side-by-side and consecutive presentations are available. In the former case, the two stimuli are presented simultaneously on the screen, and any change in viewing angle is rendered in a synchronized way, to allow users to visualize both contents from the same point of view. Conversely, in the consecutive presentation only one stimulus is presented at a time. By using the arrow keys on the keyboard, users can switch between two stimuli. The switching can happen at any viewing angle the user has chosen, thus allowing to compare the two stimuli at any point of view. A mid-grey color has been selected for the environment surrounding the simulated display, in accordance with the ITU-R recommendations [ITU-R BT.500-13, 2012]. An example rendering from the graphical interface is presented in Figure C.1.

Once users are satisfied with their viewing experience of the content, they can score the stimuli using the keyboard. All the scores are saved in an output file. The total time each stimulus was visualized is recorded in a separate file, to be used in analyzing interaction patterns and user behavior [Viola and Ebrahimi, 2017].

The software application can be found at the following link: https://github.com/mmspg/LFDisplaySimulator. It is free to use, modify or redistribute, according to the MIT license. In case of use of the software for any purpose, publishing or use of any updates and variations based on it, as well as when presenting and publishing results based on the software, a reference to [Viola et al., 2019] should be provided.

Bibliography

Edward H Adelson and James R Bergen. *The plenoptic function and the elements of early vision*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence*, 14(2):99–106, 1992.

Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafal K Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Towards a quality metric for dense light fields. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3720–3729. IEEE, 2017.

Waqas Ahmad, Roger Olsson, and Mårten Sjöström. Interpreting plenoptic images as multi-view sequences for improved compression. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

Kaan Akşit, Jan Kautz, and David Luebke. Slim near-eye display using pinhole aperture arrays. *Applied optics*, 54(11):3422–3427, 2015.

M Zeshan Alam and Bahadir K Gunturk. Deconvolution based light field extraction from a single image capture. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 420–424. IEEE, 2018.

Zahir Y Alpaslan and Hussein S El-Ghoroury. Small form factor full parallax tiled light field display. In *Stereoscopic Displays and Applications XXVI*, volume 9391, page 93910E. International Society for Optics and Photonics, 2015.

Gustavo Alves, Fernando Pereira, and Eduardo AB da Silva. Light field imaging coding: Performance assessment methodology and standards benchmarking. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016.

Jun Arai, Fumio Okano, Masahiro Kawakita, Makoto Okui, Yasuyuki Haino, Makoto Yoshimura, Masato Furuya, and Masahito Sato. Integral three-dimensional television using a 33-megapixel imaging system. *Journal of Display Technology*, 6(10):422–430, 2010.

Pekka Astola and Ioan Tabus. Wasp: Hierarchical warping, merging, and sparse prediction for light field image compression. In *2018 7th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2018a.

Pekka Astola and Ioan Tabus. Light Field Compression of HDCA Images Combining Linear Prediction and JPEG 2000. *EUSIPCO 2018*, 2018b.

Nader Bakir, Wassim Hamidouche, Olivier Déforges, Khouloud Samrouth, and Mohamad Khalil. Light field image compression based on convolutional neural networks and linear approximation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1128–1132. IEEE, 2018.

Tibor Balogh. The holovizio system. In *Electronic Imaging 2006*, pages 60550U–60550U. International Society for Optics and Photonics, 2006.

Federica Battisti, Marco Carli, and Patrick Le Callet. A study on the impact of visualization techniques on light field perception. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2155–2159. IEEE, 2018.

Bjontegaard, Gisle. Calculation of average PSNR differences between RD-curves. International Telecommunication Union, March 2001.

Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432. ACM, 2001.

Yung-Hsuan Chao, Gene Cheung, and Antonio Ortega. Pre-demosaic light field image compression using graph lifting transform. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 3240–3244. IEEE, 2017.

Shenchang Eric Chen. Quicktime vr: An image-based approach to virtual environment navigation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 29–38. ACM, 1995.

Chandrajit Choudhury, Yellamraju Tarun, Ajit Rajwade, and Subhasis Chaudhuri. Low bit-rate compression of video and light-field data using coded snapshots and learned dictionaries. In *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2015.

C. Conti, P. Nunes, and L. D. Soares. HEVC-based light field image coding with bi-predicted self-similarity compensation. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–4, July 2016. doi: 10.1109/ICMEW.2016.7574667.

Feng Dai, Jun Zhang, Yike Ma, and Yongdong Zhang. Lenselet image compression scheme based on subaperture images streaming. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4733–4737. IEEE, 2015.

Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. IEEE, Jun 2013.

Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams. Linear volumetric focus for light field cameras. *ACM Transactions on Graphics (TOG)*, 34(2), Feb. 2015.

Subbareddy Darukumalli, Peter A Kara, Attila Barsi, Maria G Martini, and Tibor Balogh. Subjective quality assessment of zooming levels and image reconstructions based on region of interest for light field displays. In *2016 International Conference on 3D Imaging (IC3D)*, 2016.

Murilo B de Carvalho, Marcio P Pereira, Gustavo Alves, Eduardo AB da Silva, Carla L Pagliari, Fernando Pereira, and Vanessa Testoni. A 4d dct-based lenslet light field codec. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 435–439. IEEE, 2018.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64 (23):6160–6173, 2016.

Touradj Ebrahimi, Siegfried Foessel, Fernando Pereira, and Peter Schelkens. Jpeg pleno: Toward an efficient representation of visual reality. *Ieee Multimedia*, 23(4):14–20, 2016.

Hussein S El-Ghoroury and Zahir Y Alpaslan. Quantum photonic imager (qpi): a new display technology and its applications. In *Invited) Proceedings of The International Display Workshops*, volume 21, 2014.

Jose N Filipe, Luis MN Tavora, Pedro AA Assuncaot, Rui Fonseca-Pinto, and Sergio MM de Faria. Evaluation of focus metrics in extended depth-of-field reconstruction. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.

Giulia Fracastoro, Dorina Thanou, and Pascal Frossard. Graph transform learning for image compression. In *Picture Coding Symposium (PCS), 2016*, pages 1–5. IEEE, 2016.

Fraunhofer Institute. Light field dataset. Available at https://www.iis.fraunhofer.de/en/ff/amm/dl/lightfielddataset.html, 2017. URL https://www.iis.fraunhofer.de/en/ff/amm/dl/lightfielddataset.html.

Qiang Fu, Zhiliang Zhou, Yan Yuan, and Bin Xiangli. Image quality evaluation of light field photography. In *IS&T/SPIE Electronic Imaging*, pages 78670F–78670F. International Society for Optics and Photonics, 2011.

Andreī Gershun. The light field. *Studies in Applied Mathematics*, 18(1-4):51–151, 1939.

Alireza Ghasemi, Nelly Afonso, and Martin Vetterli. Lcav-31: a dataset for light field object recognition. In *Computational Imaging XII*, volume 9020, page 902014. International Society for Optics and Photonics, 2014.

Bernd Girod, Chuo-Ling Chang, Prashant Ramanathan, and Xiaoqing Zhu. Light field compression using disparity-compensated lifting. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–373. IEEE, 2003.

Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54. ACM, 1996.

Danillo B Graziosi, Zahir Y Alpaslan, and Hussein S El-Ghoroury. Compression for full-parallax light field displays. In *Stereoscopic Displays and Applications XXV*, volume 9011, page 90111A. International Society for Optics and Photonics, 2014.

Petri Helin, Pekka Astola, Bhaskar Rao, and Ioan Tabus. Sparse modelling and predictive coding of subaperture images for lossless plenoptic image compression. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2016*, pages 1–4. IEEE, 2016.

Petri Helin, Pekka Astola, Bhaskar Rao, and Ioan Tabus. Minimum description length sparse modeling and region merging for lossless plenoptic image compression. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1146–1161, 2017.

Matthew Hirsch, Gordon Wetzstein, and Ramesh Raskar. A compressive light field projection system. *ACM Transactions on Graphics (TOG)*, 33(4):58, 2014.

Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Transactions on Graphics (TOG)*, 34(4):60, 2015.

ISO/IEC 29170-2. ISO/IEC 29170-2:2015 Information technology – Advanced image coding and evaluation – Part 2: Evaluation procedure for nearly lossless coding. https://www.iso.org/standard/66094.html, 2015. Accessed: 2018-07-16.

ISO/IEC JTC 1/SC29/WG1 JPEG. Grand challenge on light field image compression. Doc. M72022, Geneva, Switzerland, June 2016.

ISO/IEC JTC 1/SC29/WG1 JPEG. JPEG Pleno Call for Proposals on Light Field Coding. Doc. N74014, Geneva, Switzerland, January 2017.

ISO/IEC JTC 1/SC29/WG1 JPEG. JPEG PLENO - Light Field Coding Common Test Conditions. Doc. N80027, Berlin, Germany, July 2018a.

ISO/IEC JTC 1/SC29/WG1 JPEG. JPEG PLENO Light Field Coding VM1. Doc. N80028, Berlin, Germany, July 2018b.

ITU-T Q.6/SG 16 and ISO/IEC JTC 1/SC 29/WG 11. HEVC reference software HM. [Online]. Available: https://hevc.hhi.fraunhofer.de/trac/hevc/browser/trunk.

ITU-R BT.2022. General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays. International Telecommunication Union, August 2012.

ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, January 2012.

ITU-R BT.709-6. Parameter values for the HDTV standards for production and international programme exchange. International Telecommunication Union, June 2015.

ITU-T J.149. Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM). International Telecommunication Union, Mar. 2004.

ITU-T P.1401. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. International Telecommunication Union, July 2012.

ITU-T P.910. Subjective video quality assessment methods for multimedia applications. International Telecommunication Union, April 2008.

ITU-T P.913. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. International Telecommunication Union, March 2016.

Adrian Jacobs, Jonathan Mather, Robert Winlow, David Montgomery, Graham Jones, Morgan Willis, Martin Tillin, Lyndon Hill, Marina Khazova, Heather Stevenson, et al. 2D/3D switchable displays. *Sharp Technical Journal*, pages 15–18, 2003.

A Jagmohan, A Sehgal, and N Ahuja. Compression of lightfield rendered images using coset codes. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 1, pages 830–834. IEEE, 2003.

Changwon Jang, Kiseung Bang, Seokil Moon, Jonghyun Kim, Seungjae Lee, and Byoungho Lee. Retinal 3d: augmented reality near-eye display via pupil-tracked light field projection on retina. *ACM Transactions on Graphics (TOG)*, 36(6):190, 2017.

Adrian Jarabo, Belen Masia, Adrien Bousseau, Fabio Pellacini, and Diego Gutierrez. How do people edit light fields? *ACM Trans. Graph.*, 33(4):146–1, 2014.

Chuanmin Jia, Yekang Yang, Xinfeng Zhang, Xiang Zhang, Shiqi Wang, Shanshe Wang, and Siwei Ma. Optimized inter-view prediction based light field image compression with adaptive reconstruction. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

Xiaoran Jiang, Mikaël Le Pendu, Reuben A Farrugia, and Christine Guillemot. Light field compression with homography-based low-rank approximation. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1132–1145, 2017.

Xin Jin, Haixu Han, and Qionghai Dai. Plenoptic image coding using macropixel-based intra prediction. *IEEE Transactions on Image Processing*, 27(8):3954–3968, 2018.

Andrew Jones, Ian McDowall, Hideshi Yamada, Mark Bolas, and Paul Debevec. Rendering for an interactive 360 light field display. *ACM Transactions on Graphics (TOG)*, 26(3):40, 2007.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pages 920–929, 2016.

Peter A Kara, Maria G Martini, Peter Kovacs, Samdor Imre, Attila Barsi, Kristof Lackner, Tibor Balogh, et al. Perceived quality of angular resolution for light field displays and the validity of subjective assessment. In *2016 International Conference on 3D Imaging (IC3D)*, 2016.

Peter A Kara, Aron Cserkaszky, Attila Barst, Tamas Papp, Maria G Martini, and László Bokor. The interdependence of spatial and angular resolution in the quality of experience of light field visualization. In *3D Immersion (IC3D), 2017 International Conference on*, pages 1–8. IEEE, 2017a.

Peter A Kara, Aron Cserkaszky, Subbareddy Darukumalli, Attila Barsi, and Maria G Martini. On the edge of the seat: Reduced angular resolution of a light field cinema with fixed observer positions. In *9th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017b.

Christian Keimel, Julian Habigt, Clemens Horch, and Klaus Diepold. Qualitycrowd - a framework for crowd-based quality evaluation. In *Picture Coding Symposium (PCS), 2012*, pages 245–248. IEEE, 2012.

Yuto Kobayashi, Shu Kondo, Keita Takahashi, and Toshiaki Fujii. A 3-D display pipeline: Capture, factorize, and display the light field of a real 3-D scene. *ITE Transactions on Media Technology and Applications*, 5(3):88–95, 2017.

Koji Komatsu, Keita Takahashi, and Toshiaki Fujii. Scalable light field coding using weighted binary images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 903–907. IEEE, 2018.

Péter Tamás Kovács, Kristóf Lackner, Attila Barsi, Ákos Balázs, Atanas Boev, Robert Bregović, and Atanas Gotchev. Measurement of perceived spatial resolution in 3d light-field displays. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 768–772. IEEE, 2014.

Stanford Computer Graphics Laboratory. The (new) stanford light field archive, 2004. URL http://lightfield.stanford.edu/.

Douglas Lanman and David Luebke. Near-eye light field displays. *ACM Transactions on Graphics (TOG)*, 32(6):220, 2013.

Douglas Lanman, Matthew Hirsch, Yunhee Kim, and Ramesh Raskar. Content-adaptive parallax barriers: optimizing dual-layer 3D displays using low-rank light field factorization. In *ACM Transactions on Graphics (TOG)*, volume 29, page 163. ACM, 2010.

Douglas Lanman, Gordon Wetzstein, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar. Beyond parallax barriers: applying formal optimization methods to multilayer automultiscopic displays. In *Stereoscopic Displays and Applications XXIII*, volume 8288, page 82880A. International Society for Optics and Photonics, 2012.

Marc Levoy. Light fields and computational imaging. *IEEE Computer*, 39(8):46–55, 2006.

Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42. ACM, 1996.

Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. Light field microscopy. *ACM Transactions on Graphics (TOG)*, 25(3):924–934, 2006.

Marc Levoy, Zhengyun Zhang, and Ian McDowall. Recording and controlling the 4d light field in a microscope using microlens arrays. *Journal of microscopy*, 235(2):144–162, 2009.

Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014a.

Y. Li, R. Olsson, and M. Sjöström. Compression of unfocused plenoptic images using a displacement intra prediction. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–4, July 2016a. doi: 10.1109/ICMEW.2016.7574673.

Yun Li, Märten Sjöström, Roger Olsson, and Ulf Jennehag. Efficient intra prediction scheme for light field image compression. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 539–543. IEEE, 2014b.

Yun Li, Mårten Sjöström, Roger Olsson, and Ulf Jennehag. Scalable coding of plenoptic images by using a sparse set and disparities. *IEEE Transactions on Image Processing*, 25(1):80–91, 2016b.

Chia-Kai Liang, Tai-Hsu Lin, Bing-Yi Wong, Chi Liu, and Homer H Chen. Programmable aperture photography: multiplexed light field acquisition. In *ACM Transactions on Graphics (TOG)*, volume 27, page 55. ACM, 2008.

D. Liu, L. Wang, L. Li, Zhiwei Xiong, Feng Wu, and Wenjun Zeng. Pseudo-sequence-based light field image compression. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–4, July 2016. doi: 10.1109/ICMEW.2016.7574674.

Andrew Lumsdaine and Todor Georgiev. The focused plenoptic camera. In *Computational Photography (ICCP), 2009 IEEE International Conference on*, pages 1–8. IEEE, 2009.

Marcus Magnor and Bernd Girod. Data compression for light-field rendering. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(3):338–343, 2000.

Andrew Maimone, Gordon Wetzstein, Matthew Hirsch, Douglas Lanman, Ramesh Raskar, and Henry Fuchs. Focus 3d: Compressive accommodation display. *ACM Trans. Graph.*, 32(5): 153–1, 2013.

Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):46, 2013.

Rie Matsubara, Zahir Y Alpaslan, and Hussein S El-Ghoroury. Light field display simulation for light field quality assessment. In *SPIE/IS&T Electronic Imaging*, pages 93910G–93910G. International Society for Optics and Photonics, 2015.

Thomas Maugey, Antonio Ortega, and Pascal Frossard. Graph-based representation for multi-view image coding. *arXiv preprint arXiv:1312.6090*, 2013.

Leonard McMillan and Gary Bishop. Plenoptic modeling: An image-based rendering system. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 39–46. ACM, 1995.

Mary Meeker. Internet trends 2016 report, June 2016. URL http://www.kpcb.com/internet-trends.

Lingfei Meng, Ting Sun, Rich Kosoglow, and Kathrin Berkner. Evaluation of multispectral plenoptic camera. In *IS&T/SPIE Electronic Imaging*, pages 86600D–86600D. International Society for Optics and Photonics, 2013.

Ricardo Monteiro, Luis Lucas, Caroline Conti, Paulo Nunes, Nuno Rodrigues, Sérgio Faria, Carla Pagliari, Eduardo Silva, and Luís Ducla Soares. Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–4, 2016.

Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005.

Heinrich Niemann and Ingo Scholz. Evaluating the quality of light fields computed from hand-held camera images. *Pattern recognition letters*, 26(3):239–249, 2005.

Jens-Rainer Ohm, Gary J Sullivan, Heiko Schwarz, Thiow Keng Tan, and Thomas Wiegand. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC). *IEEE Transactions on Circuits and Systems for Video Technology*, 22 (12):1669–1684, 2012.

Roger Olsson, Marten Sjostrom, and Youzhi Xu. A combined pre-processing and h. 264-compression scheme for 3d integral images. In *2006 IEEE International Conference on Image Processing*, pages 513–516. IEEE, 2006.

Alexandre Ouazan, Peter Tamas Kovacs, Tibor Balogh, and Attila Barsi. Rendering multi-view plus depth data on light-field displays. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pages 1–4. IEEE, 2011.

Luca Palmieri, Reinhard Koch, and Ron Op Het Veld. The plenoptic 2.0 toolbox: Benchmarking of depth estimation methods for mla-based focused plenoptic cameras. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 649–653. IEEE, 2018.

Pradip Paudyal, Federica Battisti, Alessandro Neri, and Marco Carli. A study of the impact of light fields watermarking on the perceived quality of the refocused data. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*, pages 1–4. IEEE, 2015.

Pradip Paudyal, Federica Battisti, and Marco Carli. Effect of visualization techniques on subjective quality of light field images. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 196–200. IEEE, 2017a.

Pradip Paudyal, Federica Battisti, Mårten Sjöström, Roger Olsson, and Marco Carli. Towards the perceptual quality evaluation of compressed light field images. *IEEE Transactions on Broadcasting*, 63(3):507–522, 2017b.

C. Perra and P. Assuncao. High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–4, July 2016. doi: 10.1109/ICMEW.2016.7574671.

Cristian Perra. Lossless plenoptic image compression using adaptive block differential prediction. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1231–1234. IEEE, 2015.

Cristian Perra and Daniele Giusto. An analysis of hevc compression for light field image refocusing applications. In *2018 IEEE Seventh International Conference on Communications and Electronics (ICCE)*, pages 273–277. IEEE, 2018.

Cristian Perra, Wei Song, and Antonio Liotta. Effects of light field subsampling on the quality of experience in refocusing applications. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2018.

Nathanaël Perraudin, Johan Paratte, David Shuman, Lionel Martin, Vassilis Kalofolias, Pierre Vandergheynst, and David K. Hammond. GSPBOX: A toolbox for signal processing on graphs. *ArXiv e-prints*, August 2014.

Anne-Flore Perrin, Cambodge Bist, Rémi Cozot, and Touradj Ebrahimi. Measuring quality of omnidirectional high dynamic range content. In *Applications of Digital Image Processing XL*, volume 10396, page 1039613. International Society for Optics and Photonics, 2017.

Anne-Flore Nicole Marie Perrin. Context-based quality of experience in immersive multimedia. page 322, 2019. doi: 10.5075/epfl-thesis-7272. URL http://infoscience.epfl.ch/record/262805.

Tom Peterka, Robert L Kooima, Daniel J Sandin, Andrew Johnson, Jason Leigh, and Thomas A DeFanti. Advances in the dynallax solid-state dynamic parallax barrier autostereoscopic visualization display system. *IEEE transactions on visualization and computer graphics*, 14 (3):487–499, 2008.

Prashant Ramanathan and Bernd Girod. Rate-distortion analysis for light field coding and streaming. *Signal Processing: Image Communication*, 21(6):462–475, 2006.

Martin Řeřábek and Touradj Ebrahimi. New light field image dataset. 2016.

Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing (ICIP)*, pages 3097–3100. IEEE, 2011.

Mira Rizkallah, Thomas Maugey, Charles Yaacoub, and Christine Guillemot. Impact of light field compression on refocused and extended focus images. In *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016.

Mira Rizkallah, Xin Su, Thomas Maugey, and Christine Guillemot. Graph-based transforms for predictive light field compression based on super-pixels. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP 2018*, 2018.

Kunio Sakamoto and Tsutomu Morii. Multiview 3D display using parallax barrier combined with polarizer. In *Advanced Free-Space Optical Communication Techniques/Applications II and Photonic Components/Architectures for Microwave Systems and Displays*, volume 6399, page 63990R. International Society for Optics and Photonics, 2006.

Dietmar Saupe, Franz Hahn, Vlad Hosu, Igor Zingman, Masud Rana, and Shujun Li. Crowd workers proven useful: A comparative study of subjective video quality assessment. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

Ionut Schiopu and Adrian Munteanu. Macro-pixel prediction based on convolutional neural networks for lossless compression of light field images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 445–449. IEEE, 2018.

Ionut Schiopu, Moncef Gabbouj, Atanas Gotchev, and Miska M Hannuksela. Lossless compression of subaperture images using context modeling. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2017*, pages 1–4. IEEE, 2017.

Likun Shi, Shengyang Zhao, Wei Zhou, and Zhibo Chen. Perceptual evaluation of light field image. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 41–45. IEEE, 2018.

Hooman Shidanshidi, Farzad Safaei, and Wanqing Li. Objective evaluation of light field rendering methods using effective sampling density. In *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*, pages 1–6. IEEE, 2011a.

Hooman Shidanshidi, Farzad Safaei, and Wanqing Li. A quantitative approach for comparison and evaluation of light field rendering techniques. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–4. IEEE, 2011b.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Xin Su, Mira Rizkallah, Thomas Maugey, and Christine Guillemot. Graph-based light fields representation and coding using geometry information. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 4023–4027. IEEE, 2017.

Xin Su, Mira Rizkallah, Thomas Maugey, and Christine Guillemot. Rate-distortion optimized super-ray merging for light field compression. In *European Signal Processing Conference (EUSIPCO)*, 2018.

Ioan Tabus, Petri Helin, and Pekka Astola. Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and JPEG 2000. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

Keita Takahashi. Light field display project, 2018. URL http://www.fujii.nuee.nagoya-u.ac.jp/~takahasi/Research/LFDisplay/.

Keita Takahashi, Yuto Kobayashi, and Toshiaki Fujii. From focal stack to tensor light-field display. *IEEE Transactions on Image Processing*, 27(9):4571–4584, 2018.

Roopak R Tamboli, Balasubramanyam Appina, Sumohana Channappayya, and Soumya Jana. Super-multiview content with high angular resolution: 3D quality assessment on horizontal-parallax lightfield display. *Signal Processing: Image Communication*, 47:42–55, 2016.

Roopak R Tamboli, Balasubramanyam Appina, Peter A Kara, Maria G Martini, Sumohana S Channappayya, and Soumya Jana. Effect of primitive features of content on perceived quality of light field visualization. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2018a.

Roopak R Tamboli, Peter A Kara, Aron Cserkaszky, Attila Barsi, Maria G Martini, Balasubramanyam Appina, Sumohana S Channappayya, and Soumya Jana. 3d objective quality assessment of light field video frames. In *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2018b.

Evgeniy Upenik, Irene Viola, and Touradj Ebrahimi. A rendering solution to display light field in virtual reality. In *26th European Signal Processing Conference (EUSIPCO)*, number CONF, 2018.

Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69, 2007.

Ruben Verhack, Thomas Sikora, Lieven Lange, Rolf Jongebloed, Glenn Van Wallendael, and Peter Lambert. Steered mixture-of-experts for light field coding, depth estimation, and

processing. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 1183–1188. IEEE, 2017.

Alexandre Vieira, Helder Duarte, Cristian Perra, Luis Tavora, and Pedro Assuncao. Data formats for high efficiency coding of lytro-illum light fields. In *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 494–497. IEEE, 2015.

Irene Viola and Touradj Ebrahimi. A new framework for interactive quality assessment with application to light field coding. In *Applications of Digital Image Processing XL*, volume 10396, page 103961F. International Society for Optics and Photonics, 2017.

Irene Viola and Touradj Ebrahimi. Quality assessment of compression solutions for ICIP 2017 Grand Challenge on light field image coding. 2018a.

Irene Viola and Touradj Ebrahimi. Valid: Visual quality assessment for light field images dataset. In *10th International Conference on Quality of Multimedia Experience (QoMEX)*, number CONF, 2018b.

Irene Viola, Martin Rerabek, Tim Bruylants, Peter Schelkens, Fernando Pereira, and Touradj Ebrahimi. Objective and subjective evaluation of light field image compression algorithms. In *32nd Picture Coding Symposium (PCS)*, 2016a.

Irene Viola, Martin Řeřábek, and Touradj Ebrahimi. A new approach to subjectively assess quality of plenoptic content. In *SPIE Optical Engineering+ Applications*, pages 99710X–99710X. International Society for Optics and Photonics, 2016b.

Irene Viola, Martin Řeřábek, and Touradj Ebrahimi. Comparison and evaluation of light field coding approaches. *IEEE Journal of selected topics in signal processing*, 2017a.

Irene Viola, Martin Řeřábek, and Touradj Ebrahimi. Impact of interactivity on the assessment of quality of experience for light field content. In *9th International Conference on Quality of Multimedia Experience (QoMEX)*, 2017b.

Irene Viola, Keita Takahashi, Toshiaki Fujii, and Touradj Ebrahimi. A comprehensive framework for visual quality assessment of light field tensor displays. 2019.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

Gordon Wetzstein. Synthetic light field archive, 2010. URL http://web.media.mit.edu/~gordonw/SyntheticLightFields/.

Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, and Ramesh Raskar. Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting. 2012.

Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 765–776. ACM, 2005.

Jui-Yi Wu, Ping-Yen Chou, Kuei-En Peng, Yi-Pai Huang, Hsin-Hsiang Lo, Chuan-Chung Chang, and Fu-Ming Chuang. Resolution enhanced light field near eye display using e-shifting method with birefringent plate. *Journal of the Society for Information Display*, 2018.

Jason C Yang, Matthew Everett, Chris Buehler, and Leonard McMillan. A real-time distributed light field camera. *Rendering Techniques*, 2002:77–86, 2002.

Shenyang Zhao and Zhibo Chen. Light field image coding via linear approximation prior. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

Xiaoqing Zhu, Anne Aaron, and Bernd Girod. Distributed compression for large camera arrays. In *Statistical Signal Processing, 2003 IEEE Workshop on*, pages 30–33. IEEE, 2003.

Frederik Zilly, Christian Riechert, Marcus Müller, and Peter Kauff. Generation of multi-view video plus depth content using mixed narrow and wide baseline setup. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, pages 1–4. IEEE, 2012.

Irene Viola

Digital Signal Processing
and Multimedia Quality Assessment Engineer

Nationality: Italian
Address: Chemin de Montaney 44
1024 Ecublens (VD)
Email: irene.viola@epfl.ch
Mobile: +41 788355282

Research Profile

Interested in digital signal processing and multimedia compression, transmission and quality evaluation. Current research focus on compression and visual quality assessment of light field contents.

Education

- **PhD Candidate in Electrical and Electronics Engineering** — Sept. 2015 – Present
  École Polytechnique Fédérale de Lausanne (EPFL) — *Lausanne, Switzerland*
  - **Thesis**: Compression and visual quality of light field contents
  - **Advisor**: Prof. Touradj Ebrahimi

  The goal of the thesis is to provide an analysis of various methodologies for quality assessment of light field contents, to evaluate the compression capabilities of various encoding solutions, and to propose a new method to improve the coding efficiency for light field contents.

- **Master of Science in Computer Engineering** — Sept. 2013 – Oct. 2015
  Polytechnic University of Turin — *Turin, Italy*
  - **Thesis**: A dataset for image super-resolution
  - **Advisors**: Prof. Martin Vetterli, École Polytechnique Fédérale de Lausanne (EPFL)
    Prof. Enrico Magli, Polytechnic University of Turin
  - **GPA**: 3.87/4.00
  - **Grade**: 110/110

- **Bachelor of Science in Cinema and Media Engineering** — Sept. 2010 – Jul. 2013
  Polytechnic University of Turin — *Turin, Italy*
  - **GPA**: 3.69/4.00
  - **Grade**: 109/110 (2nd among graduates)

Academic activities

- **Research and Teaching Assistant** — Mar. 2016 – Present
  École Polytechnique Fédérale de Lausanne (EPFL) — *Lausanne, Switzerland*

  Research interests: Light field compression, Visual quality evaluation for immersive media
  - Active member of the JPEG standardization body. Major involvement in the JPEG Pleno group as expert in light field quality assessment. Author of 11 input documents for JPEG.
  - Leader in the European Network on Quality of Experience in Multimedia Systems and Services (Qualinet), Task Force 7 on Immersive Media Experience (IMEx).
  - Responsible assistant for Image and Video Processing, Media Security courses.
  - Main supervisor in semester projects for bachelor and master students.

- **Visiting Researcher** — Jun. 2018 – Aug. 2018
  Nagoya University — *Nagoya, Japan*

  Research interests: Evaluation of rendering-dependent compression solutions for light field displays
  Supervisor: Prof. Toshiaki Fujii

Skills

- **Programming Languages**: MATLAB, Python, Java, Bash, Javascript, C/C++, HTML/CSS
- **Softwares**: FFmpeg, Blender, LaTeX, Git, Adobe Photoshop
- **Operating systems**: Linux, MacOS, Windows

Languages

- **Full working proficiency**: English (C2)
- **Elementary proficiency**: French (A2), Spanish (A1), Greek (A1)
- **Native proficiency**: Italian

Publications

- **Journal Publications**
  - **Viola, Irene**, Martin Řeřábek, and Touradj Ebrahimi. "Comparison and evaluation of light field image coding approaches" IEEE Journal of selected topics in signal processing 11.7 (2017): 1092-1106.
- **Conference Publications**
  - **Viola, Irene**, et al. "Objective and subjective evaluation of light field image compression algorithms." Picture Coding Symposium (PCS), 2016. IEEE, 2016.
  - **Viola, Irene**, Martin Řeřábek, and Touradj Ebrahimi. "A new approach to subjectively assess quality of plenoptic content." Applications of Digital Image Processing XXXIX. Vol. 9971. International Society for Optics and Photonics, 2016.
  - **Viola, Irene**, Martin Řeřábek, and Touradj Ebrahimi. "Impact of interactivity on the assessment of quality of experience for light field content." 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017.
  - **Viola, Irene**, and Touradj Ebrahimi. "A new framework for interactive quality assessment with application to light field coding." Applications of Digital Image Processing XL. Vol. 10396. International Society for Optics and Photonics, 2017.
  - **Viola, Irene**, and Touradj Ebrahimi. "VALID: Visual quality Assessment for Light field Images Dataset." 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX).
  - **Viola, Irene**, and Touradj Ebrahimi. "Quality assessment of compression solutions for light field image coding." 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW).
  - **Viola, Irene**, and Touradj Ebrahimi. "Comparison of Interactive Subjective Methodologies for Light Field Quality Evaluation." 2018 26th European Signal Processing Conference (EUSIPCO).
  - Upenik, Evgeniy, **Irene Viola**, and Touradj Ebrahimi."A Rendering Solution to Display Light Field in Virtual Reality." 2018 26th European Signal Processing Conference (EUSIPCO).
  - **Viola, Irene**, Hermina Petric Maretic, Pascal Frossard and Touradj Ebrahimi. "A graph learning approach for light field image compression." Applications of Digital Image Processing XLI. International Society for Optics and Photonics, 2018.
  - Willelme, Alexandre, et al. "Overview of the JPEG XS core coding system subjective evaluations." Applications of Digital Image Processing XLI. International Society for Optics and Photonics, 2018
  - **Viola, Irene**, Keita Takahashi, Toshiaki Fujii and Touradj Ebrahimi. "A comprehensive framework for visual quality assessment of light field tensor displays" Electronic Imaging 2019, Society for Imaging Science and Technology (IS&T), 2019.
  - **Viola, Irene**, and Touradj Ebrahimi. "An in-depth analysis of single-image subjective quality assessment of light field contents." 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX) (pending).

Reviewer

- Conferences: International Conference on Multimedia and Expo (ICME), International Conference on Image Processing (ICIP), Data Compression Conference (DCC).
- Journals: IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Journal of Emerging and Selected Topics in Circuits and Systems, IEEE Signal Processing Letters.

Scholarships and Awards

- 2017: MKS Instruments Research Excellence Award
- 2015: EPFL EDIC Fellowship, based on outstanding academic achievements
- 2015: Full scholarship to attend European Innovation Academy (EIA)
- 2014: Erasmus+ and Swiss-European Mobility Programme scholarship
- 2010 – 2015: Full university scholarship, based on outstanding academic achievements
- 2010: Winner of the Trinity College London Olympics

Extracurricular Activities

- Amateur subtitle translator (tv comedies are my favorites!)
- Radio host for OndeQuadre's popular university radio program Zapping! (2014)

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**