



ECOLE
POLYTECHNIQUE
DE BRUXELLES

Photo-realistic depth image-based view synthesis with multi-input modality

Thesis presented by Sarah FACHADA

in fulfilment of the requirements of the PhD Degree in Engineering and
Technology (“Docteur en Sciences de l’Ingénieur et Technologie”)
Année académique 2022-2023

Supervisor: Professor Gauthier LAFRUIT
Co-supervisor: Professor Mehrdad TERATANI
Laboratory of Image Synthesis and Analysis

Thesis jury :

Olivier DEBEIR (Université libre de Bruxelles, Chair)
Dragomir MILOJEVIC (Université libre de Bruxelles, Secretary)
Maja Krivokuća (InterDigital, France)
Adrian MUNTEANU (Vrije Universiteit Brussel)
Patrice RONDAO ALFACE (Nokia)
Lu YU (Zhejiang University, China)

fnrs
LA LIBERTÉ DE CHERCHER

*“Il est essentiel de ne jamais cesser de découvrir.
En voyageant ou en restant immobile,
en échangeant ou en se taisant,
en réfléchissant ou en innovant. ”*

Pierre Bottero

Abstract

Who has never immersed themselves in the memories of a photo album or the universe rendered by a movie? Cameras have been invented to record a part of the world at a precise moment for recollection, information or entertainment purposes. However, beside the ever-increasing quality of the captured images, only one viewpoint, holding in a flat image, is captured. In order to increase the immersion, new methods have emerged to reconstruct and render the three dimensions of a scene, allowing the user to better understand, measure or simply live the captured moment.

Due to the public and industry's interest for immersive representations of the real-world (3D cinema, metaverse, virtual reality,...), the MPEG-I group (subgroup of MPEG for Immersive video coding) launched a vast work for standardization of immersive video coding. Research on view synthesis, the technique to create new viewpoints from a given set of input images, has been tackled by companies such as Google and NVIDIA, obtaining incredibly photo-realistic results. However, these methods are not yet accessible at large, due to the acquisition setups, the reconstruction process, the hardware cost or the processing time needed to render new viewpoints.

In this manuscript, we present a view synthesis method relying on the principle of depth image-based rendering: given a set of real-world images representing a scene, along with some geometry information represented by depth maps, we reconstruct new images taken from any location in the scene. We aim to reach photo-realistic results, wide navigation range, real-time processing on midrange material and multi-input modality.

We demonstrate the quality of the image synthesized by this method compared to other state-of-the-art view synthesis approaches. Thanks to a GPU implementation, we reach real-time results and high visual quality with less than ten input images. Thus, we prove that our method is suitable not only for real-time virtual reality applications, but

also for holography and 3D displays.

We also address the challenge of rendering non-Lambertian objects (e.g. mirrors, glasses, transparent liquids...). While omnipresent, they are left aside or barely mentioned in most of the view synthesis methods. Indeed, their particular interactions with light make them violate the Lambertian assumption, which is the implicit basis of most of the reconstruction and rendering methods. However, to reach a faithful rendering of the scene, and not only a plausible approximation of the scene, we extended the depth image-based rendering paradigm to reproduce their changing appearance.

Eventually, we studied the particular case of plenoptic cameras. Those devices, mimicking multi-faceted insect eyes, capture in one shot several viewpoints of the scene: exactly what is needed to overcome the gap between a flat photograph and the 3D scene in which we can immerse. However, the view synthesis with plenoptic camera, and all the related methods to make view synthesis possible (calibration, geometry estimation), are still at their first faltering steps, compared to analogous methods with regular imaging devices. As an invitation to take a journey to the plenoptic world, we explored the calibration of the plenoptic cameras, for single and multi-view datasets. We contributed to the development of view synthesis of new images within a small range. And finally, we extended our software to free viewpoint view synthesis using plenoptic cameras with calibrated parameters.

This thesis contributes to the development of photo-realistic, multi-input view synthesis. It extends the understanding of depth image-based rendering by pushing its limits beyond non-Lambertian scenes and plenoptic cameras. Furthermore, it makes the method more accessible at large through the release of open-source datasets and software, the low requirements of the proposed approaches (pattern-free calibration, portable methods) and the extensive comparisons with other state-of-the-art methods. In this sense, it opens doors to fascinating new challenges and applications in the domain of view synthesis.

Remerciements

De nombreuses personnes ont contribué à l'aboutissement de cette thèse durant les cinq dernières années et je tiens à les remercier ici.

En premier lieu, merci à mon promoteur, Pr. Gauthier Lafruit et mon co-promoteur, Pr. Mehrdad Teratani, pour leur encadrement, leur soutien, leurs conseils et leurs encouragements. Merci d'avoir cru en moi dès le début.

Merci à Pr. Olivier Debeir, Pr. Dragomir Milojevic, Pr. Maja Krivokuća, Pr. Adrian Munteanu, Dr. Patrice Randao Alface et Pr. Lu Yu qui m'ont fait l'honneur de faire partie du jury, pour leur lecture de ce manuscrit et leurs retours.

Merci au FNRS pour sa confiance en moi, merci de m'avoir permis de chercher les solutions à des problèmes qui me fascinaient depuis si longtemps. Merci au projet européen HoviTron pour son support aux projets menés à bien pendant cette thèse. Merci à Innoviris de m'avoir permis de rejoindre le projet 3DLicorneA.

Travailler au LISA, c'est aussi cotoyer des personnes extraordinaires, que ce soit face à un tableau et craie à la main, lors d'un repas, d'un séminaire ou d'un café. Merci à Pr. Christine Decaestecker et Pr. Olivier Debeir pour le soutien et leur enthousiasme. Merci à Arlette Grave pour son support administratif. Merci à Daniele Bonatto d'avoir ensoleillé mes journées à travers les nombreuses discussions que nous avons eues ensemble. Merci à Rudy Ercerk, Corentin Martens, Adrien Foucart, Thomas Vandamme, Laurie Van Bogaert, Hamed Razavi, Eline Soetens et Armand Losfeld et tant d'autres.

Merci aux professeurs qui m'ont guidée vers les sciences depuis le lycée, en particulier Madame Altschuh et Pr. Tatatiana Beliaeva.

Merci à mes amis, toujours présents malgré les kilomètres, Mercedes, Laura, Antoine, Tito, Clément, François et bien d'autres.

Merci à ma famille qui m'entoure d'amour : ma maman, Martine Dury, qui a toujours été une base arrière qui pousse en avant. Merci à Lili, ma petite *speur*, à mes enfants, Léonie et Raphaël, pour leur joyeuse contribution. Finalement, merci à mon compagnon, 'God', pour sa patience et son soutien.

Acknowledgements

Sarah Fachada is a Research Fellow of the Fonds de la Recherche Scientifique - FNRS, Belgium.

This work was supported in part by the Fonds de la Recherche Scientifique - FNRS, Belgium, under Grant n° 33679514, ColibriH.

This work was supported in part by the HoviTron project that received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N° 951989 (<https://www.hovitron.eu/>).

This work was supported in part by the FER 2021 project, Grant N° 1060H000066-FAISAN.

This work was supported in part by the Emile DEFAY 2021 project, Grant N° 4R00H000236-Emile DEFAY, Belgium.

Contents

Abstract	i
List of Figures	xi
List of Tables	xv
List of Abbreviations	xvii
1 Introduction	1
1.1 Motivation	1
1.1.1 Context	1
1.1.2 Current limitations	5
1.2 Objectives	7
1.3 Impact	7
1.4 Dissertation Overview	8
1.5 Publications	9
2 Background	15
2.1 Modeling the light field	15
2.1.1 Plenoptic function	16
2.1.2 Two planes parametrization	17
2.2 Sampling	17
2.2.1 The pinhole camera	17
2.2.2 General linear cameras	19
2.2.3 The plenoptic camera	22
2.3 Rendering	25

2.3.1	Disparity	25
2.3.2	Depth image-based rendering	26
2.3.3	Metrics for evaluation	27
3	View synthesis	31
3.1	Related work	31
3.1.1	View synthesis	32
3.1.2	Datasets	43
3.1.3	Depth estimation	45
3.2	Reference View Synthesizer	46
3.2.1	View selection	47
3.2.2	Warping	48
3.2.3	Blending	49
3.2.4	Inpainting	54
3.3	GPU acceleration for real-time navigation in virtual reality	54
3.3.1	Adaptation to the pipeline of RVS	54
3.3.2	Video player	55
3.4	View synthesis with XSlit Cameras	56
3.5	Holography	60
3.6	Evaluation	62
3.6.1	Comparison with other view synthesis methods	62
3.6.2	Impact of the depth map's quality	73
3.6.3	Speed evaluation	78
3.6.4	Holography analysis	79
3.7	Conclusions and future work	81
4	Rendering non-Lambertian objects	83
4.1	Definition and problem statement	83
4.2	Related work	85
4.2.1	Detection of non-Lambertian objects	85
4.2.2	Reconstruction	86
4.2.3	Rendering	87
4.2.4	Datasets	91
4.3	Classification	93
4.4	Processing non-Lambertian features in the 4D light field	96
4.4.1	Segmentation	96

4.4.2	Image-based rendering	97
4.5	Evaluation	100
4.5.1	Model accuracy	100
4.5.2	Comparison with other non-Lambertian view synthesis methods	103
4.5.3	Impact of the number of coefficients in the polynomial maps . .	108
4.5.4	Impact of the coefficient quantization	109
4.6	Conclusion and future work	111
5	View synthesis with Plenoptic cameras	113
5.1	Related work	114
5.1.1	Calibration	115
5.1.2	Subaperture view rendering	116
5.1.3	Datasets	119
5.2	Plenoptic Camera Calibration	120
5.2.1	Subaperture view parameters	121
5.2.2	Internal parameters	124
5.3	Rendering the subaperture images	127
5.3.1	With minimal camera parameters	127
5.3.2	With calibrated camera parameters in RVS	129
5.4	Experiments	132
5.4.1	Camera calibration	132
5.4.2	Subaperture view rendering	135
5.5	Conclusion and future work	141
6	Conclusions and prospects	143
6.1	Summary	143
6.2	Future work	144
	Bibliography	147
A	Appendix: Experiments	173
A.1	List of the used open-source datasets	173
A.2	List of the used open-source software	173
A.3	Configurations for evaluation of Chapter 3.6	174
A.3.1	Comparison with other view synthesis methods	175
A.3.2	Impact of the depth map quality	177
A.4	Configurations for evaluation of Chapter 4.5	178

CONTENTS

A.4.1	Comparison with other view synthesis methods	178
A.4.2	Impact of the number of coefficients in the polynomial maps and of the coefficient quantization	179
A.5	Configurations for evaluation of Chapter 5.4	179
A.5.1	Camera calibration	179
A.5.2	Subaperture view rendering	179
B	Appendix: Theoretical proofs	181
B.1	Feature displacement in a 2D light field	181
B.2	Proof of Equations (5.1) and (5.2)	184

List of Figures

1.1	Our rendering method running on multiple displays	3
1.2	Pipeline to render reality	4
1.3	Different types of cameras	4
2.1	The seven parameters of the plenoptic function.	16
2.2	The two planes parametrization.	17
2.3	The pinhole camera model.	18
2.4	The equirectangular projection.	19
2.5	General linear camera light sampling	20
2.6	Examples of epipolar plane images	20
2.7	The epipolar line to a point	21
2.8	Perspective in a Xslit cameras	21
2.9	The projection through a plenoptic camera.	23
2.10	The different plenoptic camera configurations	24
2.11	Disparity between two pixels	25
2.12	Principle of DIBR	26
3.1	The 4D light field.	33
3.2	Principle of light field reconstruction with Shearlet transform	34
3.3	3D reconstruction of a scene	35
3.4	Frequent artifacts in DIBR	37
3.5	Atlas representation of a scene	37
3.6	Multiplane image	40
3.7	View synthesis using LLFF	41
3.8	Training a NeRF	42

LIST OF FIGURES

3.9	Datasets for the view synthesis evaluation	44
3.10	Camera dispositions in Toystable and Fencing datasets	44
3.11	Pipeline of RVS	47
3.12	Warping phase	49
3.13	Blending phase	50
3.14	The two blending modes of RVS	50
3.15	Blending weights for background objects	52
3.16	Blending weights for foreground objects	53
3.17	The parameters of an Xslit camera.	56
3.18	Effect of changing the distance between the slits of an XSlit camera	57
3.19	Effect of changing the orientation of the slits of an XSlit camera	57
3.20	View synthesis in step-in and step-out with an XSlit camera	58
3.21	PSNR comparison for view synthesis with XSlit and pinhole inputs	59
3.22	The hogels in a holographic stereogram	60
3.23	Creation of a holographic stereogram from a sparse set of images	62
3.24	Overview of the depth maps for the view synthesis evaluation	63
3.25	Visual comparison for the Toystable dataset, 4 input images, interpolation	66
3.26	Visual comparison for the Toystable dataset, 9 input images, interpolation	67
3.27	Visual comparison for the Toystable dataset, 4 input images, step-in	68
3.28	Visual comparison for the Babyunicorn dataset, 4 input images	69
3.29	Visual comparison for the Fencing dataset, 6 input images	69
3.30	Visual comparison for the Shaman dataset, 4 and 9 input images	70
3.31	Visual comparison for the Classroom dataset, 5 input images	71
3.32	Visual comparison for the Museum dataset, 8 input images	71
3.33	Views used to compute the DERS depth maps	73
3.34	Depth maps for the evaluation of the impact of depth quality	74
3.35	Objective results of the impact of depth quality	75
3.36	Visual results of the impact of depth quality	76
3.37	Frame rate of the view synthesis using 1 to 8 input views	78
3.38	Comparison of holograms obtained with a 3D model and with DIBR	79
3.39	Comparison of holograms obtained with DIBR with 4 and 8 input images	80
3.40	Holographic stereogram of the Classroom dataset	81
4.1	Difference between perceived and ground truth depth in a mirror	84
4.2	Point triangulation is possible only in Lambertian objects	87
4.3	Disparity range for a scene with a planar mirror	89

4.4	MPI model of Nex	90
4.5	Datasets for the evaluation of view synthesis of non-Lambertian objects	92
4.6	Different EPI type observed for non-Lambertian objects	94
4.7	Segmenting a non-Lambertian object through its optical flow	97
4.8	Solving light ray correspondence for forward camera displacement . . .	99
4.9	Experiment for model validation	100
4.10	The 2D surface in the 4D light field formed by a non-Lambertian feature	101
4.11	Error maps for the polynomial approximation with degree-3 polynomials	102
4.12	Visual comparison for the Magritte Torus dataset, 25 input images . . .	105
4.13	Visual comparison for the Tarot dataset, 16 input images	106
4.14	Visual results for the impact of the model of polynomial approximation	108
4.15	Visual results for the impact of the quantization	110
5.1	Plenoptic image and conversion to multiview subaperture images	116
5.2	Conversion of a micro-image to multiview patches	117
5.3	Model of the plenoptic 2.0 camera.	121
5.4	Spatial registration of the subaperture views of a plenoptic camera . . .	122
5.5	Merging the two registrations	124
5.6	Keplerian and Galilean camera configurations	125
5.7	Flow chart to find the internal plenoptic camera parameters	127
5.8	Real and virtual images formed by the plenoptic camera	128
5.9	Pixel position in a hexagonal grid of micro-images	131
5.10	Effect of blending the results of the projections from all the micro-images.	132
5.11	Capturing condition of our multi plenoptic camera dataset	133
5.12	Output of the internal parameters' calibration for the three datasets. . . .	134
5.13	Subaperture images comparison for the Cube dataset	137
5.14	Subaperture images comparison for the Rabbitstamp dataset	138
5.15	Subaperture images comparison for the Fujita dataset	139
5.16	Subaperture images comparison for the Triview dataset	140
5.17	Projection to a plenoptic image with RVS	141
B.1	Three simple planar non-Lambertian objects	182
B.2	Baseline and rotation between two subaperture images.	184

List of Tables

3.1	Evaluated methods for view synthesis.	32
3.2	Datasets for the view synthesis evaluation	43
3.3	Objective results (MS-SSIM) for view synthesis.	64
3.4	Objective results (PSNR) for view synthesis.	64
3.5	Objective results (IV-PSNR) for view synthesis.	65
3.6	Objective results (LPIPS) for view synthesis.	65
4.1	Evaluated methods for view synthesis of non-Lambertian objects.	89
4.2	Datasets for the evaluation of view synthesis of non-Lambertian objects	92
4.3	Classification of non-Lambertian objects in function of their interaction with light.	93
4.4	Classification of non-Lambertian objects in function of their EPI.	94
4.5	EPI for some 2D non-Lambertian objects	95
4.6	Error for different subsamplings and degrees of polynomial approximation	102
4.7	Objective results for view synthesis of non-Lambertian objects.	104
4.8	Impact of the degree of the polynomial on the objective quality	108
4.9	Impact of the quantization of the coefficients on the objective quality	110
5.1	Evaluated subaperture view rendering methods.	118
5.2	Specifications of the plenoptic datasets	120
5.3	Reprojection error for each registration method	133

List of Abbreviations

- 2D** two-dimensional. 17, 34, 94–96, 100, 181
- 3D** three-dimensional. 1, 2, 5–8, 15, 22, 26, 33–36, 39, 47, 54, 55, 60, 61, 83, 85–88, 94, 96, 97, 101, 113–115, 117, 130, 131, 144
- 4D** four-dimensional. 15, 17, 21, 33, 93, 94, 97, 100
- CPU** central processing unit. 46, 55, 81
- DERS** Depth Estimation Reference Software. 45, 62, 72, 73, 77, 103, 108
- DIBR** depth image-based rendering. 5, 7, 8, 25, 26, 36–40, 46, 48, 54, 60, 72, 78–80, 83, 85, 91–93, 100, 103, 104, 108, 109, 144, 145
- EPI** epipolar plane image. 20, 21, 33, 34, 86, 88, 90, 93–95, 101, 104, 107, 144, 181
- FPS** frames per second. 55, 78, 81
- GCD** global color difference. 28
- GLC** general linear camera. 5, 7, 19, 20, 22, 56
- GPU** graphic processing unit. 39, 46, 55, 56, 63, 81, 108, 109, 143
- HMD** head-mounted display. 78
- IBR** image-based rendering. 6, 7, 35, 42, 61, 85, 87
- IV-PSNR** Immersive Video PSNR. 28, 63, 77, 174, 177
- LLFF** Local Light Field Fusion. 41, 62, 88, 104, 107, 108, 145
- LLMV** Lenslet to MultiView. 114, 116, 127, 128, 134–136, 141, 180
- LPIPS** Learned Perceptual Image Patch Similarity. 30, 63, 72, 104, 174, 177
- MIV** MPEG Immersive Video. 8, 36, 39, 144
- MLA** microlens array. 5, 22, 23, 114, 115, 120, 121, 124, 134, 135

- MPEG** Moving Picture Experts Group. 8, 121, 141
- MPEG-I** MPEG-Immersive. 8, 31, 36, 39, 43, 45, 46, 62, 81, 92, 119, 143, 144, 177
- MPI** multi-plane image. 5, 40–42, 72, 88, 90, 107, 108, 144, 145
- MS-SSIM** Multi-Scale SSIM. 29, 63, 72, 177
- MSE** mean square error. 27, 28
- NeRF** neural radiance field. 35, 42, 43, 62, 63, 72, 88, 90, 91, 103, 104, 107, 108
- NGP** neural graphics primitives. 43
- PSNR** Peak Signal to Noise Ratio. 27, 28, 63, 77, 104, 107, 177
- RDE** Reference Depth Estimation software. 45, 175, 177
- RGB** red green blue. 27, 61, 90
- RGBD** RGB+depth. 61, 79, 81
- RLC** Raytrix Lenslet Content Convertor. 8, 114, 117, 119, 127, 128, 133–136, 141, 180
- RPVC** Reference Plenoptic Virtual Camera Calibrator. 8, 179
- RVS** Reference View Synthesizer. 8, 31, 39, 46, 54–56, 62, 63, 72, 73, 77, 79–81, 83, 88, 103, 104, 107, 108, 114, 130, 135, 136, 141, 143–145, 177, 180
- SfM** structure-from-motion. 5, 6, 8, 34–36, 39, 46, 113, 121–123, 125, 133
- SSIM** Structural Similarity. 28, 29
- SVD** singular value decomposition. 98
- TMIV** Test Model of MPEG Immersive Video. 8, 36, 39, 62, 63, 72, 81, 177
- VR** virtual reality. 1, 7, 9, 32, 54
- VSRS** View Synthesis Reference Software. 8
- WVS** View Weighting Synthesizer. 39
- YUV** luminance+chrominance (blue and red). 27