

Efficient and Accurate Disparity Estimation from MLA-Based Plenoptic Cameras

M.Sc. Luca Palmieri

Dissertation
zur Erlangung des akademischen Grades
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel
eingereicht im Jahr 2020

Kiel Computer Science Series (KCSS) 2021/5 dated 2021-10-06

URN:NBN urn:nbn:de:gbv:8:1-zs-00000383-a1

ISSN 2193-6781 (print version)

ISSN 2194-6639 (electronic version)

Electronic version, updates, errata available via <https://www.informatik.uni-kiel.de/kcss>

The author can be contacted via admin@freelab.org

Published by the Department of Computer Science, Kiel University

Multimedia Information Processing Group

Please cite as:

- ▷ Palmieri L. *Efficient and Accurate Disparity Estimation from MLA-Based Plenoptic Cameras* Number 2021/5 in Kiel Computer Science Series. Department of Computer Science, 2021. Dissertation, Faculty of Engineering, Kiel University.

```
@book{Palmieri21,  
  author   = {Luca Palmieri},  
  title    = {Efficient and Accurate Disparity Estimation  
             from MLA-Based Plenoptic Cameras},  
  publisher = {Department of Computer Science, Kiel University},  
  year     = {2021},  
  number   = {2021/5},  
  doi      = {10.21941/kcss/2021/5},  
  series   = {Kiel Computer Science Series},  
  note     = {Dissertation, Faculty of Engineering,  
             Kiel University.}  
}
```

© 2021 by Luca Palmieri

About this Series

The Kiel Computer Science Series (KCSS) covers dissertations, habilitation theses, lecture notes, textbooks, surveys, collections, handbooks, etc. written at the Department of Computer Science at Kiel University. It was initiated in 2011 to support authors in the dissemination of their work in electronic and printed form, without restricting their rights to their work. The series provides a unified appearance and aims at high-quality typography. The KCSS is an open access series; all series titles are electronically available free of charge at the department's website. In addition, authors are encouraged to make printed copies available at a reasonable price, typically with a print-on-demand service.

Please visit <http://www.informatik.uni-kiel.de/kcss> for more information, for instructions how to publish in the KCSS, and for access to all existing publications.

1. Gutachter: Prof. Dr. Ing. Reinhard Koch
Christian-Albrechts-Universität zu Kiel
Kiel, Germany
2. Gutachter: Prof. Mårten Sjöström
Mid Sweden University
Sundsvall, Sweden

Datum der mündlichen Prüfung: 01-02-2021

Zusammenfassung

Der Fokus dieser Arbeit liegt auf der Verarbeitung von Bildern, die von plenoptischen Mikrolinsen-Array Kameras aufgenommen wurden. Diese Kameras ermöglichen das Einfangen eines Lichtfeldes in einer einzigen Aufnahme, wodurch mehr Informationen als bei herkömmlichen Kameras aufgezeichnet und neue Anwendungen entwickelt werden können. Jedoch führt diese erhöhte Informationsmenge zu zusätzlichen Herausforderungen und einem höheren Berechnungsaufwand. Zum einen setzt sich ein Bild aus Tausenden von Mikrolinsenbildern zusammen, was unüblich für Standard-Bildverarbeitungsalgorithmen ist. Zum anderen muss die Disparität aus diesen Mikrobildern geschätzt werden, um ein konventionelles Bild und eine dreidimensionale Darstellung zu erstellen. Daher widmet sich diese Doktorarbeit der Analyse und dem Entwurf neuer Methoden für die Bearbeitung von und Tiefenrekonstruktion auf plenoptischen Bildern.

In dieser Arbeit wurde ein Framework für plenoptische Bildern entwickelt, die alle Beiträge enthält. Eine Unschärfe-bewusste Kalibrierungsmethode für plenoptische Kameras, eine Optimierungsmethode für die Auswahl der besten Mikrolinsen-Kombinationen, eine Übersicht über verschiedene plenoptische Kameratypen und Bilddarstellungen. Datensätze, die sowohl echte als auch synthetische Bilder beinhalten, dienen zur Erarbeitung eines Richtwertes für verschiedene Disparitäts-Algorithmen und Untersuchung von Disparitätsverhalten unter verschiedenen Komprimierungsraten. Eine Methode wurde entwickelt für die Erstellung von Tiefenkarten von biologischen Mustern, die mit einem Lichtfeld-Mikroskop aufgenommen wurden.

Abstract

This manuscript focuses on the processing images from microlens-array based plenoptic cameras. These cameras enable the capturing of the light field in a single shot, recording a greater amount of information with respect to conventional cameras, allowing to develop a whole new set of applications. However, the enhanced information introduces additional challenges and results in higher computational effort. For one, the image is composed of thousand of micro-lens images, making it an unusual case for standard image processing algorithms. Secondly, the disparity information has to be estimated from those micro-images to create a conventional image and a three-dimensional representation. Therefore, the work in thesis is devoted to analyse and propose methodologies to deal with plenoptic images.

A full framework for plenoptic cameras has been built, including the contributions described in this thesis. A blur-aware calibration method to model a plenoptic camera, an optimization method to accurately select the best microlenses combination, an overview of the different types of plenoptic cameras and their representation. Datasets consisting of both real and synthetic images have been used to create a benchmark for different disparity estimation algorithm and to inspect the behaviour of disparity under different compression rates. A robust depth estimation approach has been developed for light field microscopy and image of biological samples.

Acknowledgements

I had a very good time during these years of research. I failed a lot, tried again, made a lot of errors, learned a lot of new lessons and (hopefully) made also something right. It would be nice to give back everything I got, but I guess that is a hard job, and a short acknowledgement will not do it. But it may be a start. So I want to thank everybody for their patience and compassion, two qualities I always admired. A huge thanks to my family for their infinite support, their capability of being close even when being physically far away and their teachings, to my supervisor Reinhard for his help, contribution and overall effort, to my friends who were always there, making the place where I lived my new home, to the whole team of people in Kiel and in the ETN-FPI project who had a lot of patience to correct my errors and helped my improve, made me feel part of the team and participated in several activities and to the people who have been nice with me during these years. I am not listing here everybody with their name for several reasons (maybe they are too much, or too few, or I do not want to give you this information), but I hope that everybody who helped me knows it (or at least I hope) and deserves a thanks. Also if you think you helped me, that is also enough. Thanks and keep up the good work.

Ah and lastly, thanks to you, if you are reading this. Please continue, the best is yet to come.

Luca

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Research Question	3
1.3	Potential Applications	3
1.4	Publications and Contributions	5
1.4.1	Publications	5
1.4.2	Manuscripts	6
1.4.3	Free Software Contributions	7
1.5	Structure of the Thesis	8
2	On Light Field and Plenoptic Cameras	9
2.1	State of the Art	9
2.2	How to capture the light field	11
2.3	Microlens Array based Plenoptic Cameras	14
2.3.1	Standard Plenoptic Cameras	15
2.3.2	Focused Plenoptic Cameras	17
2.3.3	Unconventional Plenoptic Cameras	22
3	Light Field Processing	25
3.1	Synthetic images	25
3.1.1	State of the Art	25
3.1.2	The Plenoptic Simulator	26
3.2	Camera Calibration	27
3.2.1	State of the Art	27
3.2.2	The Multi-focus Plenoptic Camera Model	27
3.2.3	Feature extraction	28
3.3	Analysis of the resolution and representation	31
3.3.1	State of the Art	31
3.3.2	Subaperture Images	31
3.3.3	Epipolar Plane Image	34

Contents

3.3.4	Focal stack	35
4	Disparity Estimation using Plenoptic Cameras	37
4.1	State of the Art	37
4.2	Disparity Estimation Algorithms	39
4.2.1	On microlens images	39
4.2.2	Robust Estimation for Light Field Microscopy	44
4.3	Experiments	49
4.3.1	Disparity Estimation from the Focal Stack	49
4.3.2	Disparity Estimation using Subaperture Images	52
4.3.3	In the epipolar plane image domain	54
4.3.4	Three-dimensional Reconstruction	56
4.3.5	Full three-dimensional model	57
5	Conclusions	61
6	Publications	63
6.1	Publication 1	63
6.2	Publication 2	76
6.3	Publication 3	82
6.4	Publication 4	87
6.5	Publication 5	92
7	Manuscripts	109
7.1	Manuscript 1	109
7.2	Manuscript 2	131
	Bibliography	143

List of Figures

2.1	An image captured with a plenoptic camera. Only a portion of the full image is shown here. A zoomed detail in the blue rectangle is shown to highlight the nature of a microlens array (MLA)-based plenoptic image, where thousands of microlens image (MI) are tiled together.	14
2.2	A schematic model of a standard plenoptic camera. The point p lies at the main focal plane in the scene and is projected exactly at the MLA plane, and thus seen only by one microlens. Modified image from Publication 3 in Section 6.3 [APK+18].	15
2.3	A crop of a raw plenoptic image captured with a standard plenoptic camera and its relative rendered image using [DPW13]. For a better visualization, only a portion of the scene is shown, and one detail is highlighted in the red rectangle. The raw image belongs to the dataset of Publication 3 in Section 6.3 [APK+18].	16
2.4	A schematic model of a focused plenoptic camera. A point p lies at the main focal plane in the scene and is projected between the main lens and the MLA and then imaged from several microlenses. Modified image from Publication 3 in Section 6.3 [APK+18].	18
2.5	A crop of a raw image captured with a multi-focus plenoptic camera, a subset of focused plenoptic cameras. On the right, the rendered image using the algorithm described in Section 3.3.2. For a better visualization, only a portion of the scene is shown, and two details are highlighted in the red and blue rectangle, showing how objects at different depths in the scene are imaged at different resolutions. The raw image belongs to the dataset of Publication 3 in Section 6.3 [APK+18].	19

List of Figures

2.6	The hexagonal grid of different lenses is shown in the left image. Image from [PW12]. At the right side, the a crop of a real image from the dataset captured in Publication 3 in Section 6.3 [APK+18] where the different lens types are visible.	21
2.7	Schematic model of multi-focus plenoptic camera. A point p that lies in the main focal plane of the scene is imaged onto different microlenses with different amount of blurs and magnification factor due to the different focal lengths. Modified image from Publication 3 in Section 6.3 [APK+18].	21
2.8	An early version of adobe’s light field camera’s prototype (left) and an example of a typical image captured with the prototype (right). Images from [GZC+06]	23
2.9	Design of the FiMic, the Light Field Microscope. It was designed in [SSL+18] from the research team at the University of Valencia. Image from Publication 5 in Section 6.5 [PSI+19].	24
3.1	An example of a synthetic image with its corresponding ground truth disparity. Having the ground truth disparity for each MI allows a numerical evaluation of disparity methods on the raw images. Images from Publication 2 in Section 6.2 [PKV18].	26
3.2	Results relative to the synthetic dataset. The proposed algorithm detects a larger number of corners with a lower error margin. The refinement step reduces the error at the price of decreasing the total number of detected corners. Images from Manuscript 1 in Section 7.1.	30
3.3	One example of a set of subaperture images from a scene acquired with a multi-focus plenoptic camera. The raw image belongs to the dataset acquired in Publication 3 in Section 6.3 [APK+18], while the rendering of the subaperture image (SI) is done using the algorithm described in Manuscript 2 in Section 7.2.	32

3.4	A schematic model of the rendering process. Within a microlens image, shown on the left, a grid of $s \times s$ points is used to extract a patch of size ps . That generates an image of $s \times s$ pixels, shown on the right. Modified image from Manuscript 2 in Section 7.2.	33
3.5	Epipolar plane image representation for the same scene captured with standard and multi-focus plenoptic cameras respectively. The image and their epipolar plane image representation are very similar since we tuned the parameters of the rendering, yet the slope of the lines is different because of the different sampling and the relative position of the objects in the scene. Image from Manuscript 2 in Section 7.2.	34
3.6	Changing the size of the extracted patch account for refocusing at a different depths. In the focused plenoptic camera (FPC) case, the algorithm cannot generate an all-in-focus image without knowledge about the depth of the scene, yet it can generate a focal stack of the scene, rendering images focused at different depths. Images generated using the plenoptic toolbox [Pal20] described in Publication 2 in Section 6.2 [PKV18].	36
4.1	A detail of a scene captured with a multi-focus plenoptic camera (MPC). Each color represents a different focal type. Having high virtual depth and features with high contrast is suitable for stereo matching. Image from Publication 3 in Section 6.3 [APK+18].	39
4.2	Synthetic images allow for a quantitative analysis of the effects of a similarity measures on the accuracy of the disparity estimation. <i>BadPixN</i> refers to pixels with an error higher than N , ($N = 1, 2$), <i>MSE</i> denotes the mean squared error, <i>Disc</i> and <i>Smooth</i> refer respectively to pixels belonging to depth discontinuities and to smooth regions in the scene. Modified image from Publication 2 in Section 6.2 [PKV18].	41

List of Figures

- 4.3 Different lens combinations have been evaluated and stored in a look up table to reduce the computational cost at runtime. Several combinations, along with the total number of lenses (NL) are shown in the image. The look up table shows a basic visualization of the algorithm at runtime, where combinations can be selected according to performance or accuracy. Image from Publication 1 in Section 6.1 [PK17]. 43
- 4.4 An example of an image captured with the light field microscope described in [SSL+18]. In this sample cotton fibers are imaged with a monocolored light. Image from Publication 5 in Section 6.5 [PSI+19]. 45
- 4.5 The diagram shows the workflow upon which the work was based. Several steps are used to guarantee a robust estimation for biological images. Modified image from Publication 5 in Section 6.5 [PSI+19]. 46
- 4.6 Images and respective results from different samples. It can be seen how the algorithm estimates depth with high accuracy at the micrometer scale, and a three-dimensional visualization is provided to highlight the complex structure of the scene. Samples marked with [1] were acquired in [MI15], while samples marked with [2] were captured in Valencia using the FiMic [SSL+18] as part of Publication 5 in Section 6.5 [PSI+19]. 48
- 4.7 The image shows the output of a disparity estimation using the focal stack. Because of the noisy estimation, pixels with very low confidence were filtered out, leaving a sparser disparity. This accounts for a better visualization und understanding of the regions where the estimation fails. Images generated using the plenoptic toolbox [Pal20] described in Publication 2 in Section 6.2 [PKV18]. 50

4.8 This scene is an extreme case, in which a large depth of field is required. When the foreground is sharp, the background exhibits blur similar to the gaussian one, while when the background is sharp, the foreground exhibits artefacts. This is due to the larger number of contributions from neighbouring lenses that overlap on this region. Images generated using the plenoptic toolbox [Pal20] described in Publication 2 in Section 6.2 [PKV18]. 51

4.9 An example of a scene and the disparity maps calculated with both techniques. Parameters were tuned to aim for optimal results. The iterative refinement achieves a smooth shape at the cost of losing some features of the scene, while the EPINET approach retains the structure of the scene with high fidelity although delivering an overall slightly noisier image. Original image belongs to the dataset acquired in Publication 3 in Section 6.3 [APK+18]. 53

4.10 An horizontal epipolar plane image slice from Figure 3.5. Different slopes relates to different depths of the object. The image is scaled for visualization purposes. 54

4.11 Disparity estimation from the epipolar plane image (EPI) representation. The estimation is done using the spinning parallelogram operator approach described in [ZSL+16] on images acquired in Publication 3 in Section 6.3 [APK+18]. . . 55

4.12 The scene was captured using a Raytrix R42 (MPC) and reconstructed as a point cloud based on the calibration described in Manuscript 1 in Section 7.1. 57

4.13 The setup of the experiment to capture the scene from different angles. The experiment was performed in the laboratory in University of Kiel. 58

List of Figures

- 4.14 Some results from the dataset. On the left, the point cloud created from a real image. Please notice that the background have been removed through color keying and outlier filtering. On the right, a point cloud obtained from a synthetically generated image using the simulator from Publication 4 in Section 6.4 [MPP+18]. The results on both sets look similar in terms of accuracy, with the estimation of synthetic images being slightly less noisy. 59
- 4.15 A point cloud estimated from a synthetic scene rendered using the simulator from Publication 4 in Section 6.4 [MPP+18]. The left image shows the point cloud after the reprojection from the disparity map. The noise is removed with a filter that selects the best value along the epipolar line. 60

List of Acronyms

In the text, the following acronyms are used:

MI microlens image

SI subaperture image

MLA microlens array

SPC standard plenoptic camera

FPC focused plenoptic camera

MPC multi-focus plenoptic camera

EPI epipolar plane image

Introduction

Nowadays, everyone takes pictures daily. Capturing a moment with a digital device has been incorporated into our daily routine, becoming a natural habit. Meanwhile, digital devices used to capture scenes evolved enormously from the old, large and slow machines used a couple of centuries ago to the extremely small, powerful, and quick devices of today, able to capture pictures of a quality hard to imagine just a couple of decades ago.

Recently, the resolution of the pictures acquired with high quality digital camera started to approach the limit of the human vision system. This lead research to start looking for novel possibilities to enhance the capturing process.

In fact, the fundamental problem of capturing the complex three-dimensional reality and record it on a two-dimensional sensor is still far from being solved. What we see with our eyes is the result captured on the retina of the eye from rays of light coming from different directions. In a camera, only one subset of these rays is captured and imaged onto the sensor. Because of this, not all information about the scene is recorded. An attempt at creating a realistic visualization of a complex natural scene will make it clear that the lack of geometrical information about the three-dimensional scene has to be compensated by estimating it.

The aim of computational photography is to extend the camera capabilities, incorporating additional information in the capturing process. The light field topic has shown in recent years exceptional capabilities in capturing, processing and visualizing a more comprehensive representation of our complex world.

1. Introduction

1.1 Motivation

The work presented in this thesis focus on light field processing. The research is developed within the framework of the Marie Curie European Training Network on Full Parallax Imaging, that aims at creating a network of researchers to investigate and analyze ways to record, store and manipulate more information about the light. The main objective of the project is to stimulate research towards possible ways of using novel technologies to achieve an accurate visual description of the world, from cameras capturing the light behaviour and its characteristics, to algorithms to compress and process huge trunks of recorded data and immersive displays capable of realistically recreating the three-dimensional world as we know it.

The light field denomination indicates a framework for the lights characteristics: instead of the two-dimensional conventional pictures, it aims to extend the dimensionality of the recorded information storing multiple light rays hitting the cameras, in an attempt to emulate the sophisticated human vision system.

It is well known that our vision system has an extremely complex way of combining the information recorded from each eye to create a real-time three-dimensional visualization. The light field aims at recording the necessary light rays to recreate a similar representation, particularly on the visualization of the scene from different perspectives, allowing free movement of the viewer, as in the real world. To be able to recreate a complex representation of a scene, implicit or explicit geometrical informations have to be extracted from the scene. There are several approaches for this challenge, yet some elements are constant across them.

One of the main requirements is the capturing of angular information about lights incident to the camera. A very natural operation for a person, yet a quite challenging task for a camera with a two-dimensional sensor. A formal definition of the characteristics of the lights to be captured has been introduced in [AB91] and denoted as *plenoptic function*. The name derives from *plenus*, that translates into full, and *optic*, indicating a function describing the behaviour of light in all its components.

It follows naturally that a device able to capture such function, or aiming at it, will be denoted as *plenoptic camera*. Plenoptic cameras exploits

1.2. Research Question

a clever optical configuration to simultaneously record light rays in different domains onto the same sensor. A basic representation of a camera consists of a main lens, which projects the scene directly to the sensor, where it will be captured and stored. Placing an array of microlenses between the main lens and the sensor allows to capture rays from different directions to the sensor. This accounts for trading the spatial with the angular information of the light rays. The optical configuration and its properties will be discussed in details in Chapter 2, *On Light Field and Plenoptic Cameras*.

1.2 Research Question

The core topic of this work lies in the plenoptic image processing. Capturing the scene using plenoptic cameras allows to record a huge amount of information about the light, which simultaneously makes it a very interesting field and introduces new complex challenges. Among the many challenges, the work focused on the estimation of the geometry of a scene captured with a plenoptic camera.

Many issues related to different camera models, calibrations and representation are addressed in Chapter 3, *Light Field Processing*, creating a suitable framework for this thesis and future works. The core of this work lies in the estimation of the geometry of the scene, investigating the optimal working pipeline for different types of images and MLA-based plenoptic cameras and evaluating different disparity estimation approaches, and is discussed in Chapter 4, *Disparity Estimation using Plenoptic Cameras*.

1.3 Potential Applications

Although light field and plenoptic cameras have been only recently an active research topic, many applications have been explored and showed promising results.

The first plenoptic cameras available for the consumer market was the Lytro [Lyt18], creating a handheld plenoptic camera able to capture the light field. Despite its broad diffusion, the company moved to virtual reality and eventually closed, discontinuing the production of those cameras.

1. Introduction

Because of its ease of use, this camera was widely used in many academic projects, constituting the basis for many light field related research. An overview of the most important is given in Chapter 2, *State of the Art and Related Works*.

Plenoptic cameras are currently produced by Raytrix [Ray20], which aims more at industrial applications. Their camera has been shown to deliver high quality results in tasks such as three-dimensional monitoring, robotics, particle flow tracking and microscopy. Exciting works have also shown promising results for engineering tasks. For example, this technology enabled three-dimensional internal visualization of an engine during combustion, providing a powerful tool for optical diagnostics [CLS16]. It has also been applied to material recognition, proving a clear advantage with respect to conventional cameras. In [WZH+16] a dataset has been created ad-hoc for this task. An adequate use of the informations recorded from these camera also allows for capturing and recovering the structure of objects through different medium, making possible underwater imaging even in low light conditions [MKR+17; LLU+18].

Outside of engineering, biology and microscopy have proven to be exciting and interesting fields for these technologies, where the angular and particularly the depth information delivers a great advantage for many analysis. Several works have been proposed in the microscopy field [LNA+06; LZM09; MI15; SSL+18; SPS+19], yet to the best of our knowledge, apart from Raytrix cameras, no commercial implementations are available at the moment of the publication. For example, experiments showed promising results in volumetric recording of fluorescent activity, which is an essential tool in modern microscopy. A plenoptic camera allowed volumetric reconstruction of the entire probe in real-time based on a single light-field recording in [SB17]. The applications of such a technology are not strictly related to microscopy. Promising results for reconstruction of biological samples create hope for future research in this direction. In [BSH+17] a light field otoscope is developed, able to recover the first three-dimensional reconstructions of children tympanic membranes in normal and otitis media conditions.

The light field is in continuous development, and new types of devices are continuously appearing. The patent from the Lytro company was acquired by Google [Goo20], who produced their own rig for cap-

1.4. Publications and Contributions

turing the light field [OEE+18], consisting of a large amount of cameras tiled together in an arc-shaped arm to capture panoramic light fields. To perform an actual breakthrough and be available the consumer market, plenoptic cameras still need to be improved. Most of the technology used to have high requirements, making it more suitable for academic research than consumer application. However, in the last years commercial implementations proved interesting in this field: Nokia [Nok20] distributes a smartphone with 5 Zeiss cameras on the back, Light [Lig20] created a camera prototype that combines 16 different cameras with different optics and LG was granted a patent for a smartphone with 16 cameras arranged as a two-dimensional array on the back [KLK+19; KLK+18].

1.4 Publications and Contributions

A part of the work on this thesis consisted in the creation of a free software framework to work with plenoptic cameras. More specifically, the work addressed particularly multi-focus plenoptic cameras. Softwares and datasets are important to promote research in this field, and at the beginning of the project very little material was available, motivating the effort in this direction.

1.4.1 Publications

During the research period, the following articles were accepted and published:

1. *Optimizing the lens selection process for multi-focus plenoptic cameras and numerical evaluation*, Luca Palmieri and Reinhard Koch, published in the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops in 2017, where the lens selection process is investigated thoroughly, providing a deeper knowledge about the selection of the best combinations of lenses to be chosen in the disparity estimation,
2. *The Plenoptic 2.0 Toolbox: Benchmarking of Depth Estimation Methods for MLA-Based Focused Plenoptic Cameras*, Luca Palmieri, Ron op het Veld

1. Introduction

and Reinhard Koch, published in the 25th IEEE International Conference on Image Processing (ICIP) in 2018, where the topic of depth estimation and the challenge about its quantification is assessed and the first version of the plenoptic toolbox has been released, consisting of both working code and a dataset including both real and synthetic images.

3. *Matching Light Field Datasets From Plenoptic Cameras 1.0 And 2.0*, Waqas Ahmad, Luca Palmieri, Reinhard Koch and Mårten Sjöström, published in the 2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), where the difference between the two commercial implementations of standard and focused plenoptic cameras, respectively Lytro and Raytrix, are discussed and the first dataset consisting of the same scene captured with both cameras is made available.
4. *Simulation of plenoptic cameras*, Tim Michels, Arne Petersen, Luca Palmieri and Reinhard Koch, published in the 2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), in which the rendering process for synthetic images that simulates real plenoptic cameras is investigated, and an open-source version of the simulator is made available.
5. *Robust Depth Estimation for Light Field Microscopy*, Luca Palmieri, Gabriele Scrofani, Nicolò Incardona, Genaro Saavedra, Manuel Martínez-Corral and Reinhard Koch, published in Sensors in 2019, in which a novel solution for the depth estimation of biological structure using images captured using the Fourier Integral Microscopy is proposed.

1.4.2 Manuscripts

Moreover, the work includes two more manuscripts. One was submitted:

6. *Geometric Calibration of Multi-Focus Plenoptic Cameras*, Nuno Barroso Monteiro, Luca Palmieri, Tim Michels, Leandro Cruz, Reinhard Koch, Nuno Gonçalves, José António Gaspar, submitted to IEEE Transactions on Cybernetics in May 2020, where a full geometric calibration is

1.4. Publications and Contributions

proposed, from an automatic detection and extractions of the corners, to the parameters estimation including the blur on different microlens types.

And one is in preparation:

7. *An Overview of Plenoptic Cameras: from Disparity to Compression*, Luca Palmieri, Waqas Ahmad, Mårten Sjöström and Reinhard Koch, where the disparity task is evaluated over different plenoptic representation and estimation methods, and the behaviour of plenoptic images and their disparity under different compression rates is analyzed.

1.4.3 Free Software Contributions

Moreover, during this research period, I actively developed or participated to the following free softwares and datasets:

- ▷ The Plenoptic Toolbox 2.0 [Pal20]: a python framework to work with plenoptic 2.0 images, where the main functions to work with those images are already implemented. Functionalities as reading and writing of images, disparity estimation, view rendering, refocusing and 3D visualization are provided, and the code is open for future extensions.
- ▷ Plenoptic Datasets: because of the computational and economical efforts that plenoptic 2.0 camera require, having access to high quality images becomes a hard task. Guided by this considerations, two different datasets consisting of scenes of various nature and suitable for different purposes were generated:
 1. Plenoptic 2.0 Dataset [Pala]: this dataset consists of images from the multi-focus plenoptic cameras. It contains images captured with two different cameras, both from the Raytrix company, namely the R29 and R42, the highest quality camera available at the time of the publication, plus some images rendered with the initial version of the plenoptic simulator. The synthetic rendered images are available with the corresponding disparity ground truth.

1. Introduction

2. Matching Plenoptic 1.0 and 2.0 [AP]: this dataset aims specifically at the research focusing on difference and similarities of different types of plenoptic cameras. The best available commercial implementation of both camera types were selected, respectively Raytrix R29 and Lytro Illum at the time of the publication, and the same scene was captured using both cameras.
- ▷ Robust Depth Estimation for Light Field Microscopy [Palb]: the software to calculate depth maps from images acquired using a light field microscopy has been released for comparison and extensions. The code is easily adaptable to different microscopes or camera configurations.
 - ▷ Plenoptic Simulator [MPP+]: a Blender plug-in to realistically emulate the rendering process for plenoptic images has been released to enable further research.

1.5 Structure of the Thesis

This thesis is organized as follows: after the first chapter gives an introduction to the topic and illustrates the research question and the motivation behind the work, the second chapter gives an overview of the theory behind plenoptic cameras and their working principles.

The third chapter focuses on plenoptic image processing, from the generation of synthetic images, to the camera calibration and their representations. The core of this work is included in the fourth chapter, where the disparity estimation is analyzed and discussed, along with experiments performed during the research period.

Conclusions are drawn in the fifth chapter and the last two chapters enlist publications and manuscripts resulted from this thesis.

Instead of having a whole chapter about the related work and the state of the art in the field, each chapter has its own related state of the art section for an easier understanding of the topic.

On Light Field and Plenoptic Cameras

This chapter gives an overview about the concept of light field and plenoptic cameras. Once we described the plenoptic function, we summarize the most used approaches to sample it. Although this work focuses on plenoptic cameras, literature shows there are other similar devices which are worth mentioning.

2.1 State of the Art

The literature about light field spans more than a century. Since the first mention in [Gab08] light field related technologies evolved significantly, showing a number of diverse approaches in attempt to capture the plenoptic function. Thanks to the recent developments, these technologies achieved in this field results that were unthinkable some decades ago. As a result, light field has become an active topic of research and several devices to capture the light field has been deployed.

Pioneering work in this direction has been the investigation of the plenoptic function and the elements of early vision in [AB91], that laid the basis for the successive works on capturing the light field with a single lens [AW92] and the reparameterization and rendering of the recorded information [GGS+96; LH96].

Possibly, the first milestone for the plenoptic camera diffusion was the introduction of the handheld model [NLM+05]. Thanks to the work on the optical design and the software needed for capturing and rendering, the camera allowed to capture the light field in a single shot with a modified

2. On Light Field and Plenoptic Cameras

version of a conventional camera. This helped significantly its diffusion and promoted research on captured light fields. This camera model placed a MLA between the main lens and the sensor, at the focal plane of the main lens. They are denoted as standard plenoptic camera (SPC) and their geometry has been extensively studied [HAH+14a; HAH+14b; HAV+16].

The major drawback of these cameras is the spatial resolution. Consequently, the discussion leaned towards the understanding of the different trade-offs between angular and spatial resolution that control the capturing process when using such cameras [LFD08] and the different optical configuration that could allow a more efficient sampling of the plenoptic function [GZC+06].

This, in turns led to the development of the FPC, described in [LG09]. The difference with respect to the above mentioned SPC consisted in changing the MLA position to modify the sampling of the plenoptic function. This change affects the angular and spatial resolution of the captured and rendered image. As analyzed in [GZC+06], an increase in one domain results in a reduction in the other one.

If for the case of SPC the rendering process relies solely on the raw data and the camera calibration [DPW13], in the case of FPC the rendering algorithm needs knowledge about the scene to render the whole scene sharply. Extracting spatial information by selecting one pixel behind each microlens is applicable in the SPC case, where every pixel behind in the MIs capture the same scene point from a slightly different perspective and accordingly the spatial information of these pixels is redundant [DPW13], yet leads to artifacts for the FPC when different objects appear at different depths in the scene [LG09]. This has been taken into account in a successive work [GL10], which introduced the idea of incorporating depth information or the knowledge about the geometry of the scene in the rendering algorithm.

The image rendering process for FPC and the achievable spatial resolution were widely discussed, for example in [WFJ11] a gradient-based approach is used to determine the size of the patch of the spatial information to be extracted from each microlens, without directly involving a depth estimation. The range of the spatial resolution was tested against their theoretical model in [DOS+14], confirming the insight that objects at different depths in the scene are imaged at different resolution, introducing

2.2. How to capture the light field

additional challenges in the image rendering process.

A special case of the FPC, denoted as MPC, was introduced in [GL12; PW12] to increase the depth-of-field of plenoptic cameras. As pointed out in [GL12], a small aperture is needed to have a larger depth of field. Yet, for the case of plenoptic cameras, a larger aperture is needed to pair the main lens with the MLA, introducing a hard constrain on the depth of field. However, using different focal lengths, different parts of the scene can be focused simultaneously, allowing, through a dedicated rendering algorithm, to extend the depth of field for plenoptic images.

Because of the optical configuration needed to capture information using a MLA, plenoptic camera usually exhibit narrow field of view. A recent work proposed to build a plenoptic camera with wide field of view [DSF+17]. The idea is to use a mono-centric lens to collect panoramic-like plenoptic images. Fisheye lenses, that could achieve a similar result in terms of field of view, suffer from a fundamental problem in the entrance pupil, whose diameter is too small and reduces the depth sensitivity. As a proof of concept, a system in which a LF camera is rotated around a monocentric lens was built and a panoramic image is captured and processed to correct for aberrations.

2.2 How to capture the light field

As mentioned, the field of computational photography aims at recording any visual information about the light in the scene. Describing the light behaviour is a complex modelling task. In the computational photography field the most common description refers to the plenoptic function.

The plenoptic function was first introduced in [AB91] as the seven-dimensional function described in Equation 2.2.1. Although the function aims to describe the full spectrum of the light, it is clear that the information that can be captured with nowadays technologies is only a subset of it.

$$P = P(\theta, \phi, \lambda, t, V_x, V_y, V_z) \quad (2.2.1)$$

The distribution of light intensity P depends on the three-dimensional spatial position (V_x, V_y, V_z) , the angle of the incident light rays (θ, ϕ) , the



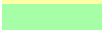
2. On Light Field and Plenoptic Cameras

wavelength of the light (λ) and its time instant (t).

Ignoring the light spectrum and the direction of the incident light rays, a conventional camera just captures a three-dimensional slice of the light field.

Table 2.1. Table with light field acquisition methods. It uses and updates data collected in [HSJ+14].

Type	LT	SS	SD	RAW	SPA	ANG	COM
Standard Plenoptic Camera	High	Yes	Yes	High	Low	High	Low
Focused Plenoptic Camera	High	Yes	Yes	High	Mid	Mid	Mid
Pinhole Masks	Low	Yes	Yes	Low	Low	Low	Low
Coded Aperture	Mid	Yes	Yes	Low	Low	Low	Mid
Scanning Pinhole	Low	No	Yes	High	High	Mid	Low
Camera Array	High	Yes	No	High	High	Mid	Mid
Compressive LF	Mid	Yes	Yes	High	High	Mid	High
Angle Sensitive Pixels	High	Yes	Yes	Mid	Mid	Mid	Mid

Color	Meaning	Acronym	Meaning
		LT	Light Transmission
		SS	Single Shot
		SD	Single Device
		RAW	Raw Image Resolution
		SPA	Spatial Resolution
		ANG	Angular Resolution
		COM	Computational Effort
	Sub-Optimal		
	Interesting		
	Optimal		

Different technologies were proposed to sample the light field. No device manages to capture the whole plenoptic function, yet they manage to capture spatial and angular information of the light in the scene.

Although there are several common criteria to select different optical

2.2. How to capture the light field

configuration to sample the light field, trade-off between characteristics of the lights are mostly application depending, posing a complex challenge in the search for an optimal solution. The data collected in [HSJ+14] about light field capturing technologies is extended in Table 2.1 with updated data, dividing plenoptic cameras into two subcategories for a more clear picture and choosing some different criteria for the columns.

To simplify the classification, values were divided into *Sub-Optimal*, *Interesting* and *Optimal*. Intuitively, the approaches containing a larger number of *Optimal* labels are more promising. The idea behind labelling a field as *Interesting* is to focus on different software approaches that can be used to extract information from a specific device. It is interesting to note that a more efficient approach could result in higher resolution without changing the capturing device.

In this analysis technologies which do not allow capturing in a single shot or using a single device are considered severely limited, making them less appealing for most purposes. Because of this, we do not investigate further Scanning Pinholes and Camera Arrays, even though the latter are widespread in the literature and have several advantages for capturing static scenes, which are still the standard, due to the fact plenoptic videos constitute still a complex challenge for processing pipelines.

Pinhole Masks have the advantage of little effort in the hardware creation, yet they suffer from limitation in terms of light transmission and resolution of the rendered image [VRA+07]. A more modern approach in [MUG18], uses a random color coded mask to perform different shots maximizing the incoherency of the measurement matrix and allowing the recording of dynamic scenes. Using a coded aperture or mask is very similar to the latter. The coding improves the light transmission while requiring a higher computational effort for the rendering process [LFD+07]. A pretty recent approach improved the spatial resolution by using coded aperture cameras [IKT+18]. A neural network is trained as an auto-encoder for plenoptic images to improve the final rendered image. This allows to decrease the requirements on spatial and angular density in the capturing process. Compressive light field approaches constitutes a solution for reaching good resolution in both spatial and angular domains, yet they sacrifice light transmission and have higher requirements in terms of computational effort for processing and rendering [MWB+13].

2. On Light Field and Plenoptic Cameras

Angle sensitive pixels were introduced in [SWG+11] and used in [HSJ+14] to capture sparser light fields that, after processing, can be converted into high resolution images. Yet the linear reconstruction achieves only low quality and the non-linear is computationally expensive, introducing a hard challenge to solve [HSJ+14].

2.3 Microlens Array based Plenoptic Cameras

This section is focused on MLA-based plenoptic cameras, exploring more in details how these cameras sample and represent the light field. As previously mentioned, the characteristic of MLA-based plenoptic camera allows to capture simultaneously light in the spatial and angular domains.

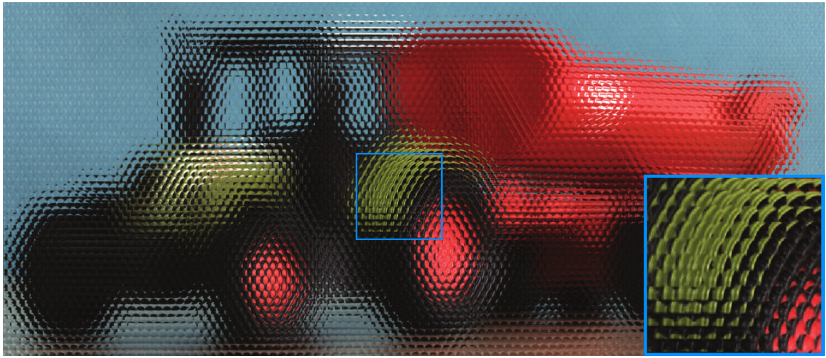


Figure 2.1. An image captured with a plenoptic camera. Only a portion of the full image is shown here. A zoomed detail in the blue rectangle is shown to highlight the nature of a MLA-based plenoptic image, where thousands of MI are tiled together.

The MLA is placed in the camera body, between the main lens and the sensor. This allows the microlenses to split light rays coming from different directions so that they can be captured separately by the sensor. In Figure 2.1 an example image is shown. It is possible to notice how one plenoptic image is formed by thousands of small MI tiled together. In each of these small MI a different piece of the information is stored. The sampling of the light field is regulated through the optical configuration

2.3. Microlens Array based Plenoptic Cameras

of the plenoptic camera, which includes main lens, MLA and the sensor. The most common MLA-based plenoptic camera models are introduced in the following sections, to have a deeper understanding of the sampling process for each model.

2.3.1 Standard Plenoptic Cameras

The standard plenoptic camera is the first type of handheld plenoptic camera that was proposed. Introduced in [NLM+05], its design consists in placing the MLA exactly at the focal plane of the main lens, as visible in Figure 2.2.

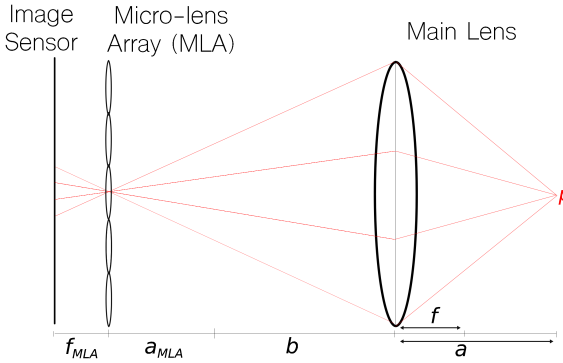


Figure 2.2. A schematic model of a standard plenoptic camera. The point p lies at the main focal plane in the scene and is projected exactly at the MLA plane, and thus seen only by one microlens. Modified image from Publication 3 in Section 6.3 [APK+18].

This configuration account for particular characteristics. As shown in Figure 2.2, to study the behaviour of the light rays one can apply the thin lens equation $\frac{1}{f} = \frac{1}{a} + \frac{1}{b}$ [Hec98] to the above schematic model, where f denotes the focal length, a the object distance, between the object and the lens, and b the image distance, between the lens and the projection point. f_{MLA} and a_{MLA} refer to the MLA distances. It follows that a point p that lies on the focal plane of the main lens, is imaged to its conjugate plane directly on the MLA plane.

2. On Light Field and Plenoptic Cameras

Looking closer at one single microlens, its imaging process is unconventional, because the point is projected exactly at the MLA plane. This translates into a blurred image on the sensor. However, the information on the sensor can be transformed to a more intuitive representation. In fact, each different pixel behind a microlens represents the angular information of one point. This means that it records the intensity of the light rays coming from the scene from different directions, and each pixel can be considered as the discretised version of the information coming from a different angle.

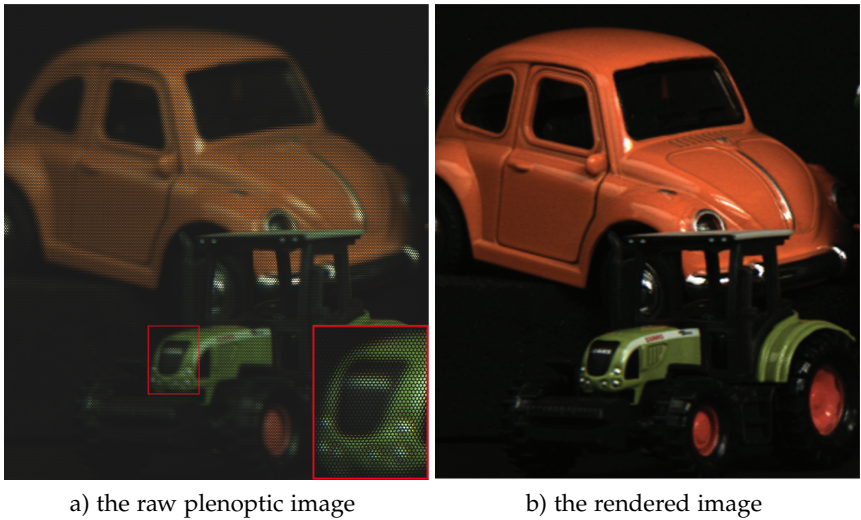


Figure 2.3. A crop of a raw plenoptic image captured with a standard plenoptic camera and its relative rendered image using [DPW13]. For a better visualization, only a portion of the scene is shown, and one detail is highlighted in the red rectangle. The raw image belongs to the dataset of Publication 3 in Section 6.3 [APK+18].

An example of a picture taken with a standard plenoptic camera is shown in Figure 2.3. On the left, the raw image is visible. The image looks blurred and contours appear not well-defined due to the presence of the MLA. In fact, the left image is just the collection of all the MI captured from each micro-lens. On the right, a conventional image was rendered from

2.3. Microlens Array based Plenoptic Cameras

the captured scene using the toolbox in [DPW13]. The rendering algorithm consists in selecting one pixel from each MI formed behind a microlens. By selecting the central pixel in each MI, a conventional image can be formed. This is denoted as central view.

Behind each microlens there are several pixels, thus different perspective views can be extracted. If instead of selecting the central pixel, a shift is applied and a pixel in the same relative position is selected from each MI, the image is rendered from a different perspective. In this work these images are denoted as SI. The different pixels behind a single microlens account for the creation of parallax between views, both in the horizontal and vertical direction. The number of pixels behind a microlens corresponds to the number of viewpoint images that can be rendered from a scene without interpolating new data.

An important trade-off in designing such cameras is the spatial and angular resolution, which intrinsically depends on the microlens size. Larger microlenses or higher sensor resolution result in more pixels behind a single microlens. Intuitively, the more pixels behind a microlens, the larger the angular resolution and the number of views that can be rendered. On the other hand, the spatial resolution is controlled from the number of microlenses in the MLA. Thus, selecting one pixel per microlens, the final spatial resolution of a rendered view corresponds exactly to the number of microlenses. Therefore smaller microlenses lead to higher spatial resolution. This mechanism is well-known as spatio-angular resolution trade-off and has been discussed in [GZC+06]. Because of commercialization by Lytro [Lyt18] and their low computational effort, SPC faced a wide spread and have been widely used in the research field.

Currently the interest in the SPC has decreased, on one side because of the lower spatial resolution of the rendered images, on the other side because of the discontinuation of the main producer of such cameras, namely the Lytro [Lyt18] company.

2.3.2 Focused Plenoptic Cameras

The second plenoptic camera model, denoted as focused plenoptic camera, was developed to address the limitations of the SPC, mainly the spatial resolution and the depth of field. As investigated in [GZC+06], an increase

2. On Light Field and Plenoptic Cameras

in the spatial resolution lead to a reduction of the resolution in the angular domain. Since most of the applications do not require high angular density and novel view estimation shows promising results with plenoptic images, such trade-off is considered beneficial.

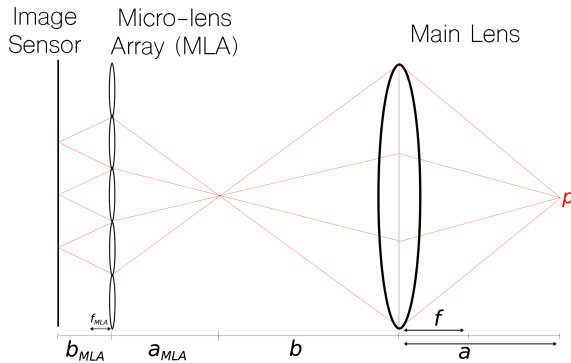


Figure 2.4. A schematic model of a focused plenoptic camera. A point p lies at the main focal plane in the scene and is projected between the main lens and the MLA and then imaged from several microlenses. Modified image from Publication 3 in Section 6.3 [APK+18].

The imaging process for plenoptic cameras depends on the optical configuration, thus on the placement of the MLA: in particular, the distances sensor to MLA and MLA to the main lens controls the mentioned trade-off. In the case of the SPC, as explained above, the projection of a real world point p lying on the focal plane of the main lens lands exactly on the MLA plane. In the focused case, the projection through the main lens lands on a distinct plane, creating a different scenario. The projected point is sharply imaged from several microlenses that see a portion of the scene, as shown in Figure 2.4, where f , a and b are respectively focal length, object distance and image distance as explained in the previous section and follow the thin lens equation [Hec98]. If we look solely at the microlens image formation process, we see that the MLA act as a camera array. Because of its narrow field of view, however, each microlens records only a small portion of the scene. It is important to notice that points at different depths will

2.3. Microlens Array based Plenoptic Cameras

be reprojected at different distances before or behind the sensor, thus resulting in different magnification factors and amount of blur.

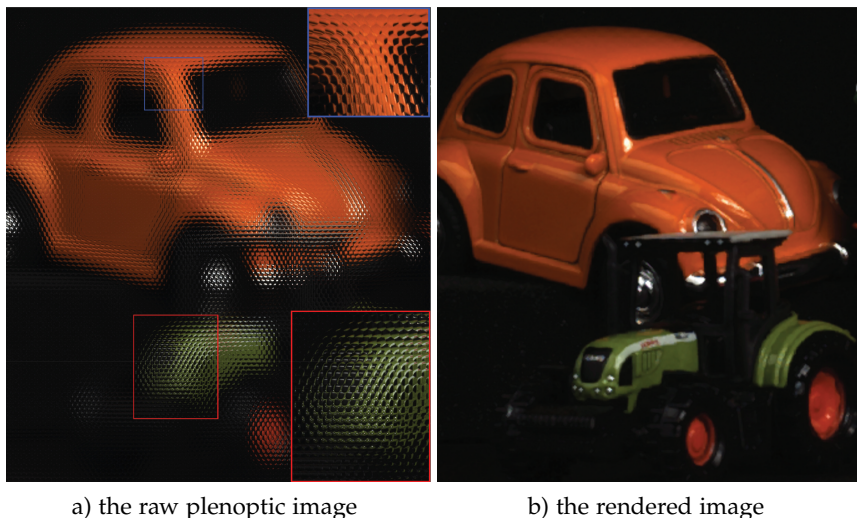


Figure 2.5. A crop of a raw image captured with a multi-focus plenoptic camera, a subset of focused plenoptic cameras. On the right, the rendered image using the algorithm described in Section 3.3.2. For a better visualization, only a portion of the scene is shown, and two details are highlighted in the red and blue rectangle, showing how objects at different depths in the scene are imaged at different resolutions. The raw image belongs to the dataset of Publication 3 in Section 6.3 [APK+18].

There are two possible scenarios, depending on the relationship between the focal length of the main lens and the scene. The first, denoted as the Galilean configuration, happens when the scene projected through the main lens, called the virtual intermediate image, lies between the MLA and the main lens. Because of its similarity with conventional cameras, this case is the most intuitive and easier to visualize, therefore is the one depicted in Figure 2.4. As for the case of standard cameras, the images created on the sensor are inverted in this configuration.

The second scenario, that goes under the name of Keplerian configuration, originates when the virtual intermediate image lands behind the

2. On Light Field and Plenoptic Cameras

sensor, making it harder to visualize. In this case, the scene will undergo a double projection and each MI records an image without flipping. Even though the virtual intermediate images changes its location, the two configurations are equivalent up to a flipping of the images on their plane, so the analysis can be done in either cases.

In Figure 2.5 we show one case of a real image acquired using a R29 Raytrix Camera where the optical setup is configured following the Keplerian model. That means that the virtual image projected through the main lens lies behind the sensor.

Here the above mentioned effect is clearly visible. Notice how a point closer to the camera is reprojected behind the sensor further away and thus imaged onto multiple microlenses, as the object shown in the red rectangle. It follows that the object has a lower spatial resolution and a larger angular resolution. The opposite is true for the object in the blue rectangle, which being further away to the camera is imaged closer to the sensor, resulting into higher spatial and lower angular resolutions. This illustrates the changes in resolution and the challenge in detecting and predicting how many times one object will be imaged onto the sensor.

Multi-Focus Plenoptic Cameras

Even though there are many possible optical configurations for plenoptic cameras, some parameters suffer from hard constraints. In order to correctly couple the focal length of the main lens with the one of the microlenses, a large aperture is needed. Having a small aperture translates in low light transmission and affects the MI quality. The disadvantage of a large aperture is the narrow depth of field, and FPC constitutes no exception. To overcome this issue and improve the depth of field of the scene, a novel technique was introduced in [GL12], where a new MLA consisting of an array of interleaved different microlenses was proposed. By using two different focal lengths, each microlens type focus at a different distance, thus doubling the depth of field when the images from different microlenses are merged together.

This idea was extended in [PW12]. Instead of two different microlens types, three lens types with three different focal length were arranged in an hexagonal pattern, as in Figure 2.6. This allowed to extend the depth of

2.3. Microlens Array based Plenoptic Cameras

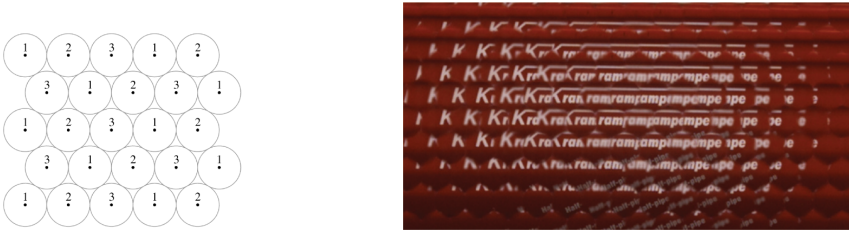


Figure 2.6. The hexagonal grid of different lenses is shown in the left image. Image from [PW12]. At the right side, the a crop of a real image from the dataset captured in Publication 3 in Section 6.3 [APK+18] where the different lens types are visible.

field of a factor of 3 and accounted for a loss in the spatial resolution of the rendered image by a factor of 2.

Looking at the image formation model of these cameras, one can notice different conditions with respect to the previous FPC case. For an easier and more coherent explanation, again the Galilean configuration is selected and drawn in Figure 2.7. The virtual intermediate image then lies between the MLA and the main lens.

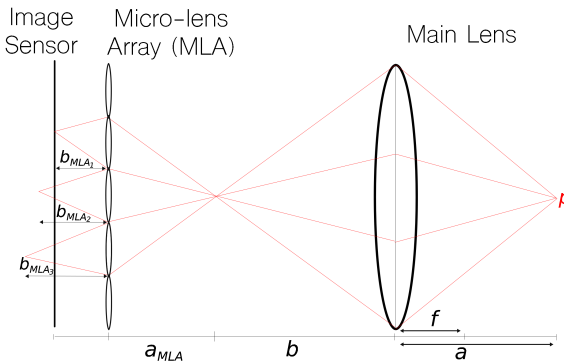


Figure 2.7. Schematic model of multi-focus plenoptic camera. A point p that lies in the main focal plane of the scene is imaged onto different microlenses with different amount of blurs and magnification factor due to the different focal lengths. Modified image from Publication 3 in Section 6.3 [APK+18].

2. On Light Field and Plenoptic Cameras

The difference in this case is the focal lengths of each microlens. A real world point projected through the main lens is imaged differently through each microlens type. A point sharply recorded onto the sensor through one particular microlens implies it is recorded with a different magnification factor and some degree of blur from the other two microlens types.

This introduces additional challenges in the rendering step, since only focused microlenses should be selected to reconstruct a sharp image. Moreover, as above mentioned, the spatial resolution is reduced, since the number of focused lenses is a fraction of the number of microlenses in the MLA. Nevertheless, the actual trade-off between spatial resolution and depth of field has proven to be effective and cameras with this configuration are still used in research and industry. At the moment of the publications, these cameras are the only plenoptic cameras still commercially produced and sold by Raytrix [Ray20].

2.3.3 Unconventional Plenoptic Cameras

The two plenoptic cameras described above are the most widely used in academic research and industrial applications. However, other optical configurations were developed and tested for different applications. Some of them are worth mentioning as a contribution to the discussion, creating an interesting perspective about different attempts at sampling the light field, yet most of them are discontinued or unique prototypes.

Adobe's Prototype

The first prototype for handheld plenoptic camera was developed from Adobe and comprised of a small array of 19 negative lenses, capturing different perspective of the scene.

The camera was built in order to capture the whole scene in each lens, therefore each lens is bigger than the microlenses in the previously seen plenoptic cameras. The actual capturing in this case results in a series of SIs directly imaged to the sensor. These images are equivalent to the standard plenoptic camera raw images. The two representations can be obtained one from the other, and the spatio-angular trade-off is regulated through the choice of the optical configuration.

2.3. Microlens Array based Plenoptic Cameras



Figure 2.8. An early version of adobe's light field camera's prototype (left) and an example of a typical image captured with the prototype (right). Images from [GZC+06]

These prototype was not continued and did not make it to the market as a commercial implementation. Nevertheless, the idea of recording the sub-aperture images directly onto the sensor has its own advantages, as it reduces the pre-processing to get the rendered images.

Light Field Microscopy

Microscopy is a very interesting field for light field technologies, since angular and depth information can lead to dramatic improvement and may be more valuable than higher spatial resolution. The actual possibility of having real-time light field recording on a microscope, enabling the estimation of the three-dimensional geometry of biological scenes is a very promising application and the reason behind several attempts in this field.

Different optical configuration have been proposed, starting from the light field microscopy project at Stanford [LNA+06], then brought on and expanded with several contributions [BGY+13; CYA+14] and later further developed by other research group [SPS+19] or slightly modified to enhance the resolution of the captured sample [SSL+18].

Although different in the implementation, these works share the same concept, using the information recorded in the light field to reconstruct the three-dimensional shape of the inspected scenes. Recently the technology of light field microscopy is reaching a wider audience and first prototypes are starting to appear, as it is the case for the Doitplenoptic [doi20] company, a university spin-off based on the work described in

2. On Light Field and Plenoptic Cameras

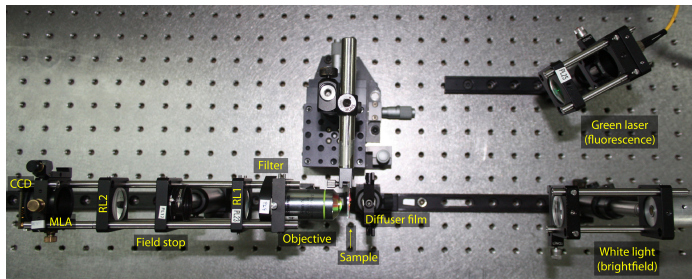


Figure 2.9. Design of the FiMic, the Light Field Microscope. It was designed in [SSL+18] from the research team at the University of Valencia. Image from Publication 5 in Section 6.5 [PSI+19].

[SSL+18] using the FiMic shown in Figure 2.9. Having recorded light field in real time, such information can be used to estimate a depth map for biological sample, a challenge that was addressed in this thesis during the collaboration with the University of Valencia, whose result is available in Publication 5 in Section 6.5 [PSI+19]. The depth estimation approach is also discussed in Section 4.2.2.

Light Field Processing

This chapter covers the processing on plenoptic images. The topics of generation of synthetic images, calibration and the different representations and the resolution trade-off are addressed and discussed in the following.

3.1 Synthetic images

To enhance research on plenoptic cameras, the creation of accurate and realistic synthetic images is a powerful resource. Because of the complicated optics of these cameras, however, the approaches proposed in the literature still show limitations. One of the main advantages of having synthetic images is the possibility of creating a ground truth image for the disparity or depth estimation, to be used for evaluation of different techniques.

3.1.1 State of the Art

Some attempts have been made to provide a light field benchmark, achieving only a partial solution. The first light field benchmark was introduced in [WMG13], consisting of dense sets of subaperture images. In this case, the dataset consisted of both synthetic and real images. Synthetic images were rendered using the Blender [Ble20] engine, while real images were acquired using a gantry to move around the camera and a light scanner to capture the ground truth. A similar dataset was proposed also in [HJK+16], although here only synthetic images are available. This resulted in an effort to evaluate all available methods for disparity estimation for the light field. However, no information about MLA is taken into account. Images are rendered as if they were taken from a moving camera or an

3. Light Field Processing

array of cameras. Due to this approach, no algorithm dealing with MI can be included in the benchmark.

3.1.2 The Plenoptic Simulator

This motivates our work in Publication 2 in Section 6.2 [PKV18], where the first available ground truth data per MI is made available online for benchmarking purposes. These images do not take into account main lens aberrations and consider microlenses as ideal lenses. This step allowed the generation of the ground truth disparity.

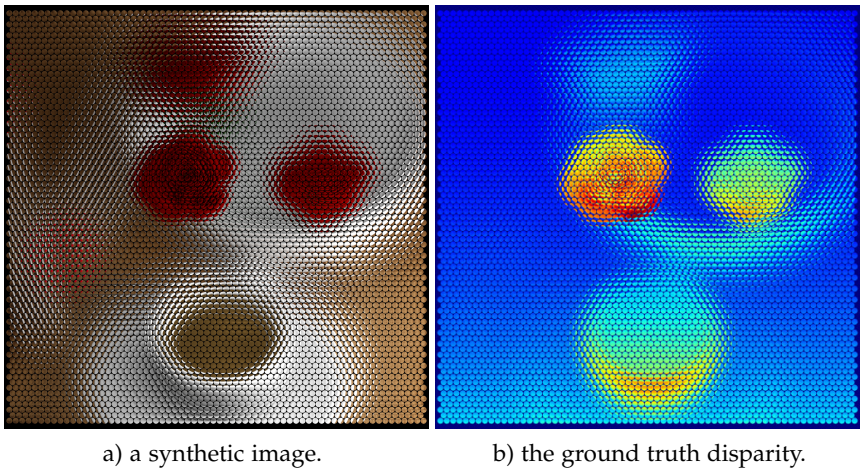


Figure 3.1. An example of a synthetic image with its corresponding ground truth disparity. Having the ground truth disparity for each MI allows a numerical evaluation of disparity methods on the raw images. Images from Publication 2 in Section 6.2 [PKV18].

An effort into creating more realistic and physically accurate results has been made in Publication 4 in Section 6.4 [MPP+18], where all optical parts of a plenoptic camera have been re-created using the Blender [Ble20] simulation engine. If on one side this leads to more realistic results, on the other end it does not account for the creation of ground truth disparity, leaving margin for future works.

3.2 Camera Calibration

Having a camera model able to describe the whole range of plenoptic cameras is a hard challenge, and most of the approaches focus on one specific type.

3.2.1 State of the Art

Although plenoptic calibration has been studied in the literature, most methods starts from the SPC case and extend to FPC, but fails to account for the MPC, including only one lens type and neglecting the analysis of the blur.

One common approach for calibration of SPC was published along with a light field toolbox in [DPW13] and it maps rays defined in pixels (i, j) and microlenses (k, l) indices to rays defined by a position (s, t) and a direction (u, v) in metric units by means of a 5×5 H matrix. This was extended in [BJK17] where a novel geometric calibration using line features is proposed. Line feature are not the only possibility to detect features on small MI, and the detection and extraction of corners through the analysis of circular boundaries is discussed in [BHK16]. Different methods have been proposed to include the FPC in the calibration. In [JHG+13] a dedicated target on a linear axis is used to detect dots and measure the ground truth distance to compare with the results from the calibration. An automated methods that accounts for the MPC is proposed in [HSH+16] and constitutes the core of the software used by the Raytrix [Ray20] company. A geometrical approach using line through the region of maximum intensity has shown accurate results in estimating the optical parameter of a MPC in [NCP+17]. Finally, an approach able to generalise the light field parameterization accounting for SPC as well for FPC has been investigated in [ZZL+18] achieving high quality results using a multi-projection-center model with 6 intrinsic parameters.

3.2.2 The Multi-focus Plenoptic Camera Model

A camera model has been developed to take into account the MPC case and its blur levels. The model describes the projection of a world point

3. Light Field Processing

through the camera optics with an affine mapping consisting of 7 intrinsic parameters and includes $M + 1$ additional parameters (where M is the number of microlens types) to model the blur in the MIs.

The calibration process is based on the automatic detection, clustering and extraction of corner points and blur radius from the raw images, therefore working on MIs. To ensure robustness to noise, an algorithm for detecting features was developed combining local and epipolar-based approaches, described in details in the next section.

Both synthetic and real images were used to evaluate the calibration against state of the art techniques, showing accurate results in the detection of the corners, the estimation of the blur radius and the estimation of the optical configuration of the camera through the calibration process.

More details about the calibration and the affine mapping are available in Manuscript 1 in Section 7.1, while a deeper discussion about the feature extraction follows next.

3.2.3 Feature extraction

As mentioned, the calibration relies on the extraction of features from raw images captured with MPC. Because of the physical implementation of the MLA, in the raw image between each MI there is a small dark area. These regions introduce additional noise and false matches on MI edges, and thus must be avoided or removed. Therefore, before extracting the feature, the MLA grid is detected and used to avoid the dark areas and detect the features in each MI.

In fact, the biggest difference with respect to conventional corners detection algorithm consists in the nature of the images. Instead of having an image of the scene, the detection works on a smaller MI which contains only a part of the scene. One MI has a diameter that ranges from 15 to 40 pixels for most common commercial implementations, and exhibits different degrees of defocus in the MPC case, where each microlens type has a different focal length. These conditions are sub-optimal for standard corner detectors, that usually fails to correctly detect the corners position [NCP+17].

Moreover, while in conventional images each corner appears only once, in the MPC case each corner is recorded several times, and no prior

3.2. Camera Calibration

knowledge on the amount can be assumed, excluding a coarse estimation of the upper and lower bound related to camera optics and scene size.

A novel method was proposed in Manuscript 1 in Section 7.1 based on these considerations. The idea for the feature extraction comes from merging different approaches from literature, where the problem of estimating corners in MI was addressed. An important distinction has to be made based on the plenoptic camera model: if in SPC or FPC the MI are focused at the same depth plane, in the MPC case the MI show different amount of blur. The proposed approach is calibrated to take into account the blur and incorporate into the theoretical camera model.

Since MPC are still a niche topic and most calibration methods only addressed SPC or FPC, there are no more than a few proposed solutions. The most interesting two are the following: in [BHK16], circular boundaries are used to estimate corners of a checkerboard. In this case avoiding to analyze the central pixel ensure the robustness against the noise. On the other side, their approach is based on detecting a sharp change between black and white region, that is not the case for blurred images. In [NCP+17] another strategy is applied, using the region of maximum intensity to create lines whose intersect reveals then the exact corner location.

The developed method consists in an attempt of extracting the advantages of both techniques and overcoming their limitations. It consists of four main steps:

1. As an initial step, a combined strategy is used: the analysis of the circular boundaries resemble the one described [BHK16] to detect the ideal shape, yet instead of looking at the gradient between white and dark color, the search is concentrated at the regions of maximum intensity as in [NCP+17], detecting the four areas which create the typical shape of a corner. It calculates the similarity to the ideal shape for each boundary of each pixel of each MI, creating a likelihood map. Since this is a computationally expensive operation, a simple thresholding approach is applied to each MI to filter out the one clearly without corner.
2. Since the same checkerboard corner appears in more than one MI, a clustering step is performed at this time. Creating clusters not only helps in the further refinement, but also to remove outliers.

3. Light Field Processing

3. Within each cluster we refine the corners positions estimation using epipolar line geometry. The assumption is that corners within a single cluster should lie on the same epipolar line. Corners that do not comply with the assumption are eliminated, reducing the total number and increasing the overall accuracy.
4. For each corner we estimate the blur radius in pixel. The additional information is incorporated into the camera model for a more accurate reconstruction. A focus measure from the literature [PPG13], the Tenengrad Variance, have been chosen and adapted to our specific purposes.

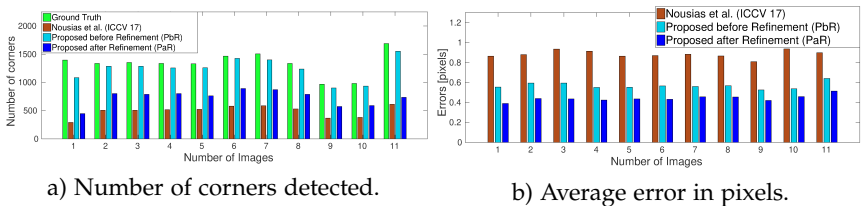


Figure 3.2. Results relative to the synthetic dataset. The proposed algorithm detects a larger number of corners with a lower error margin. The refinement step reduces the error at the price of decreasing the total number of detected corners. Images from Manuscript 1 in Section 7.1.

The algorithm was tested on both synthetic and real data for a quantitative evaluation and was shown to outperform state-of-the-art on plenoptic images, allowing for a more accurate calibration. In Figure 3.2 numerical results based on synthetic images with ground truth are reported. For more results on real image and on the calibration accuracy, please refer to Manuscript 1 in Section 7.1. Moreover, with respect to state-of-the-art method, the proposed algorithm incorporates the blur information, which is usually neglected.

3.3 Analysis of the resolution and representation

As discussed above, the term *plenoptic camera* refers to a broad range of devices with different characteristics. This motivates the need for a unified framework and a standard representation for the light field captured from different devices. Since the actual information is higher dimensional, the conventional two-dimensional image is not the most suitable representation. A unified framework allows different approaches to be applied to any type of plenoptic capture. However, several challenges arise because of the fundamental limitations in the spatio-angular resolution.

3.3.1 State of the Art

Plenoptic cameras capture higher dimensional information onto a two-dimensional sensor, therefore a suitable representation is needed to describe how the information is arranged. Although the EPI representation, a *three-dimensional description of a static scene from a dense sequence of images*, was introduced first in [BBM87], the two plane parameterization described in [GG96] became the most widely used parameterization for the plenoptic function [AB91]. Another approach [DPW13] investigated and implemented the pipeline to transform raw images into SI and is widely used as basis for further research. A more recent work [MRI17] describes any light field camera as an Equivalent Camera Array (ECA) aiming to build a unified model. The spatio-angular trade-off and its characteristics based on the plenoptic camera model are discussed in [GZC+06], which states that an increase in one domain will result in a decrease in the other domain.

3.3.2 Subaperture Images

One way to represent the light field is to create a collection of images taken from different viewpoints, emulating the capturing process of a moving camera. The process consists in reorganizing the pixel structure, extracting information from the MIs and combining into a multi-dimensional array of SIs.

3. Light Field Processing

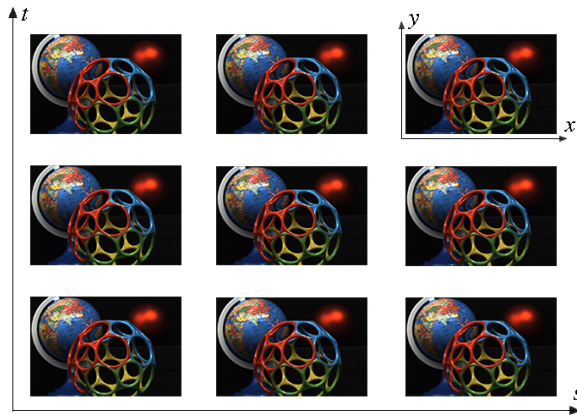


Figure 3.3. One example of a set of subaperture images from a scene acquired with a multi-focus plenoptic camera. The raw image belongs to the dataset acquired in Publication 3 in Section 6.3 [APK+18], while the rendering of the SI is done using the algorithm described in Manuscript 2 in Section 7.2.

By using the two-plane parameterization from the lumigraph [GGS+96], the light field can be depicted as in Figure 3.3. The parameterization describes the light field with 4 parameters, with (x, y) indicating the spatial resolution of a viewpoint image and (s, t) corresponding to the angular resolution. Each camera model has its own advantages and limitations regarding the resolution, and most are application depending.

For the case of SPC, a transformation can be applied from MIs to SIS without prior knowledge of the scene, relying on the camera calibration. Because of the different configuration, the rendering process for the MPC case requires some prior knowledge about the scene. The rendering consists in tiling together patches extracted from each MI, and the size of the patch depends on the scene and cannot be known beforehand. Example of solutions are available in [GL10], where depth information is used to select the correct patch size and in [WFJ11], where the correct value of the patch size is found by analysing the gradients at the border between MI for different patch sizes.

In Manuscript 2 in Section 7.2, an improved version of the render-

3.3. Analysis of the resolution and representation

ing algorithm for MPC is proposed, which allows more flexibility when rendering SIS. It was used to render the SIS in this work, as in Figure 3.3.

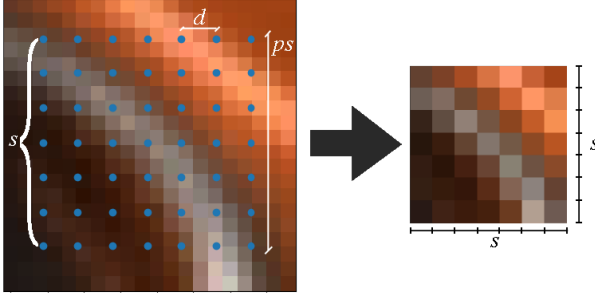


Figure 3.4. A schematic model of the rendering process. Within a microlens image, shown on the left, a grid of $s \times s$ points is used to extract a patch of size ps . That generates an image of $s \times s$ pixels, shown on the right. Modified image from Manuscript 2 in Section 7.2.

The rendering algorithm is based on tiling together patches extracted from each MI. Instead of extracting the patches as a rectangle of pixels, a grid of points is used, as shown in Figure 3.4. The patch size ps is related to the virtual depth of the point and to the MLA calibration, as investigated in [GL10; PW12]. That implies that the algorithm requires the estimation of a disparity map per MI. However, since for each MI one patch has to be extracted, only one value of disparity per MI is needed. The single value can be estimated as a weighted average or with more sophisticated methods, ensuring robustness against noisy estimations.

The number of sample per lens, denoted in Figure 3.4 with the letter s , controls the spatial resolution of the rendered image. A small value of s will result in down-scaling most of the patches, while larger values s will up-scale them. As to be expected, larger images are more prone to exhibit artefacts due to interpolation. The interpolation is done using bivariate spline approximation over a rectangular mesh, which is computationally expensive, yet more accurate. The distance between two points in the sampling grid d is derived from these two parameters, as $d = \frac{ps}{s-1}$.

The extraction of sub-pixel information allows to handle with higher precision the different patch size and the scaling factor. This is the main

3. Light Field Processing

difference with previous approaches and it accounts for several advantages: the spatial resolution can be regulated through the s parameter, and the scaling of the MI is incorporated. Moreover, when generating multiple views, the parallax between views can be controlled with sub-pixel accuracy, allowing for the generation of sparser or denser sets of SIs, that plays a determinant role in the formation of sparser or denser light field representations.

3.3.3 Epipolar Plane Image

As mentioned above, EPIs were initially introduced as a description for dense image sequences, where the camera movements between different acquired frame is relatively small. The distance between different view points is depending on the optical setup, yet these conditions are satisfied in light field capturing using most MLA-based plenoptic cameras.

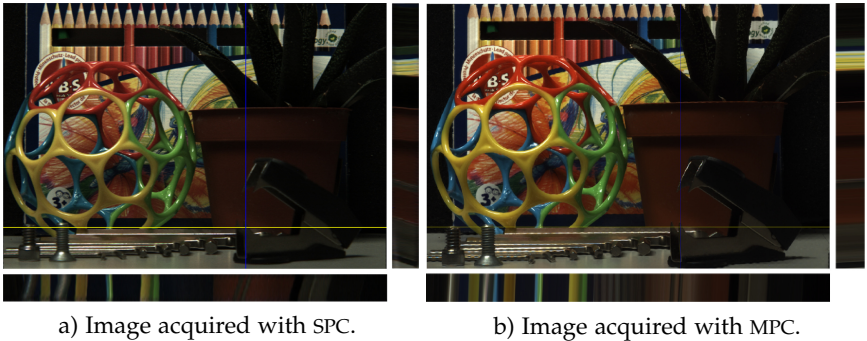


Figure 3.5. Epipolar plane image representation for the same scene captured with standard and multi-focus plenoptic cameras respectively. The image and their epipolar plane image representation are very similar since we tuned the parameters of the rendering, yet the slope of the lines is different because of the different sampling and the relative position of the objects in the scene. Image from Manuscript 2 in Section 7.2.

The EPI representation stores the information captured from the light field in a 4-dimensional structure. The advantage of this structure is the possibility to extract 2-dimensional slices from different axis, both from

3.3. Analysis of the resolution and representation

the spatial and angular domain.

In a conventional two-dimensional image, one would obtain a single line by fixing one axis and traversing the image on the other one. In the EPI however, one can traverse the four-dimensional volume along another direction and obtain a two-dimensional slice. Extracting two-dimensional slices enhances the possibility to analyse the relationship between changes in two dimensions. More in details, a point visible in one SI is transformed into a line in an EPI slice. The slope of the line is directly proportional to the three-dimensional position of an object in the scene with respect to the camera coordinates, its depth. The slope relates to the parallax shift of one point between views. Intuitively, objects further away from the camera will exhibit more parallax, thus their slope will have stronger inclination. For plenoptic cameras, one needs to take into account that the optical configuration consists of two lenses. One way to see this is to look at the intermediate image after being projected through the main lens. The closer the projected point is to the MLA, the larger the parallax and the disparity and the smaller the angle of its line.

In Figure 3.5 an example of two-dimensional slices extracted from a light field captured with both SPC and MPC is shown next to the SI. A characteristic of these slices is the limited resolution, since the range of each axis corresponds to the number of views in that direction. Extracting accurate information from these slices requires both high angular resolution and density. A high angular resolution is needed to have enough views to render meaningful lines. A high angular density enforces rendering smooth lines, whereas low angular density may break the lines in separate segments. It follows that the EPI representation is more suitable for light fields captured with SPC, that exhibits larger angular resolution. Nevertheless, because of the redundancy in the light field, the above mentioned rendering algorithm is able to generate a set of images with large angular density from a scene captured MPC using the information from the disparity estimation.

3.3.4 Focal stack

Another possibility is to rearrange the information recorded in the light field to generate a focal stack of the scene, creating a set of images focused

3. Light Field Processing

at different focal planes within the scene range. In this case, usually only a subset of the actual information is actually exploited, and the scene is captured from a single perspective, although it is possible to extend it to create a focal stack for each SI generated. The focal stack is generated directly from the MIs by extracting and tiling together patch of different sizes, and do not need any depth information beforehand. However, the focal stack stores informations about the depth of the scene and allows for more applications, and is therefore an interesting case study.



a) A scene with the background object in focus.

b) Same scene with focus on the foreground object.

Figure 3.6. Changing the size of the extracted patch account for refocusing at a different depths. In the FPC case, the algorithm cannot generate an all-in-focus image without knowledge about the depth of the scene, yet it can generate a focal stack of the scene, rendering images focused at different depths. Images generated using the plenoptic toolbox [Pal20] described in Publication 2 in Section 6.2 [PKV18].

To render a sharp image acquired with a FPC or a MPC, it is important to select the adequate patch size. The patch size is related to the depth of an object in the scene. Therefore, by gradually changing the size of the extracted patch, the rendered images will be focused at different planes, as visible in Figure 3.6. Even though for the rendering of these images no prior information about the geometry of the scene is required, for the particular case of the MPC the knowledge of the different lens types helps in seamlessly fusing together the patches extracted from different microlenses avoiding artefacts.

Disparity Estimation using Plenoptic Cameras

As discussed in the previous chapter, one of the most important and challenging topics in the plenoptic imaging is the disparity estimation. Additionally, in the FPC case the information about the geometry of the scene is needed to render all-in-focus images and thus in almost all light field powered applications. In this chapter different methods for estimating disparity from plenoptic images are discussed. The analysis about the disparity estimation is conducted on both real and synthetic images, where ground truth is available. The topic has been investigated in two of the publications contained in this thesis, and is discussed in Section 4.2.

4.1 State of the Art

Disparity estimation is based on the analysis of different visual cues and different representations of the light field. Because of the large amount of data captured in the light field and its intrinsic redundancy, several approaches can be applied.

A common approach consists in creating a cost volume in order to extract a disparity map. This is the case for techniques that operate directly on the raw plenoptic images. In [FK14] the cost volume is built first and then a regularization step is applied to calculate a disparity map for each microlens. The work was further developed in Publication 2 in Section 6.2 [PKV18] analyzing the cost function and the similarity measures using as a benchmark a set of synthetic images. Stereo correspondences can also be directly reprojected into a three-dimensional point cloud, as done in

4. Disparity Estimation using Plenoptic Cameras

[FG16], yet the accuracy of the results do not differ significantly with the respect to the disparity calculated on the raw plenoptic images.

The limitation of these approaches lies in the noisiness of the cost volume, that prevent the extraction of the true disparity in regions poor of textures. In an attempt of overcoming this challenge, focus based estimation methods have been proposed. In [LCB+15] two properties of the focal stack are exploited: the assumption that non-occluding pixels exhibit symmetry along the depth dimension and the data consistency between the synthesized image and the acquired one. Computing the occlusion matte and depths of thin structures from a focal image stack was done by modelling the spatially variant blur and the mutual occlusions at different depths in [LND17]. In [HSV+17] the focal stack is used to create stereo pairs of images. Computing their disparity allows to calculate the position of a point in the scene. Introducing partial focal stacks to deal with occlusions and increase robustness yielded more accurate results in the disparity estimation. A recent work [SAG17] links depth to normal estimation through a regularization step and a joint optimization, outperforming previous works.

Another approach consists in combining different cues to achieve a higher accuracy and less noisy results. This idea was used in [THM+13] where correspondence and defocus cues were fused and extended in [TSM+15] by adding also a shading component. In [JPC+15] an initial distortion correction is applied before building the cost volume and an iterative refinement process is used to optimize disparity. Another possibility to estimate the disparity is the analysis of angular patches extracted from the light field, as in [CLY+14; MBG18]. This idea was further developed in [WER15] to handle occlusions.

A recent work used a fully-convolutional neural network to achieve fast and accurate light field depth estimation by considering the light field geometry and proposing light field specific data augmentation methods [SJY+18]. Although limited synthetic data was available for the training, results are quite accurate.

Lastly, epipolar plane image representation has been largely used for disparity estimation of light fields. Analyzing the ray space of four-dimensional light field is the common approach in this case. Modelled as a variational multi-label problem to be solved with global optimization

[WSG13], EPI-based disparity methods achieved high quality results. This can be done exploring the geometric structures of its three-dimensional lines for triangulation and stereo matching [YGL+13] or integrating a spinning parallelogram operator into a depth estimation framework to remove the influence of occlusions [ZSL+16].

4.2 Disparity Estimation Algorithms

In this chapter the development of disparity estimation algorithms is investigated, whose related contributions consist in the algorithm to estimate disparity from the MIs, optimizing the selection process, creating a benchmark for comparisons and developing a robust approach to compute depth maps on images acquired with a light field microscope. Finally, promising experiments on disparity estimation and three-dimensional reconstruction from disparity maps have been conducted using different input representations, as EPI or the rendered focal stack.

4.2.1 On microlens images

Working on images acquired with FPC and MPC, the transformation from MI to SI requires information about the scene geometry, i.e. the disparity.

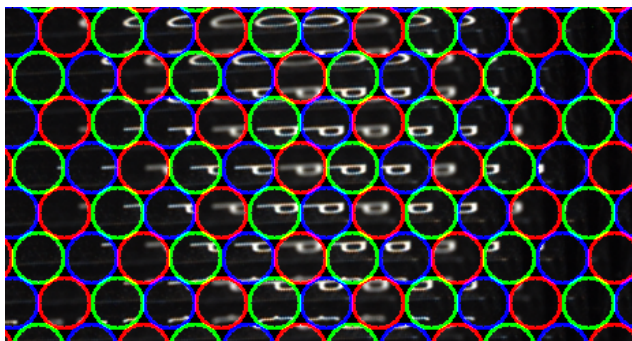


Figure 4.1. A detail of a scene captured with a MPC. Each color represents a different focal type. Having high virtual depth and features with high contrast is suitable for stereo matching. Image from Publication 3 in Section 6.3 [APK+18].

4. Disparity Estimation using Plenoptic Cameras

A suitable approach for estimating the disparity in such images is to perform the operations directly on the MIs. This technique, followed in Publication 1 (Section 6.1, [PK17]) and Publication 2 (Section 6.2, [PKV18]), consists in applying stereo matching on neighbouring microlenses. Stereo matching does not require a global structure and works with local structures. Results using sparse [FG16] and dense [FK14; PKV18] estimation achieved similar quality.

The proposed approach is based on the extraction of the minimum labels from a cost volume. The cost volume is built by calculating the similarity for each pixel in each MI along the epipolar lines, and is defined as:

$$C_{reg}(\mathbf{p}, d, \mathcal{L}_{ref}) = \frac{1}{|\mathcal{N}|} \sum_{\mathcal{L}_n \in \mathcal{N}} \mathcal{F}(\mathbf{p}, d, \mathcal{L}_{ref}, \mathcal{L}_n) \quad (4.2.1)$$

Where we denote the cost \mathcal{C} relative to a pixel p , a disparity label d and the reference lens \mathcal{L}_{ref} as the sum of the costs of the similarity function \mathcal{F} for each of the MI \mathcal{L}_n belonging to the neighbourhood \mathcal{N} . The factor $\frac{1}{|\mathcal{N}|}$ is used for normalization.

The objective is to have a cost volume robust to noise, therefore an appropriate similarity measure is selected. The largest problem in matching based on pixel similarity are regions poor of textures. In fact, it is very hard to find reliable correspondences in these regions. Based on the literature [HS08], an analysis of the most promising similarity measures has been made to select the most suitable one.

Having carried out the experiments on synthetic images, the results are quantitative. The full analysis is reported in Publication 2 in Section 6.2 [PKV18], here a brief overview is given. Even though outcomes of different measures are comparable, two methods achieve the highest accuracy overall. They are the sum of absolute value (SAD) and the CENSUS measures, which have more robust performances across different measurements criteria. Normalized cross correlation (NCC) behaves interestingly, because even though it performs above average in particular criteria as depth discontinuities (*BadPixNDisc* in Figure 4.2), it exhibits large errors in smoother regions (*BadPixNSmooth* in Figure 4.2). In this case, a combined solution could be able to limit errors while exploiting features of different methods. Sum of squared values (SSD) and gradient (GRAD) have shown

4.2. Disparity Estimation Algorithms

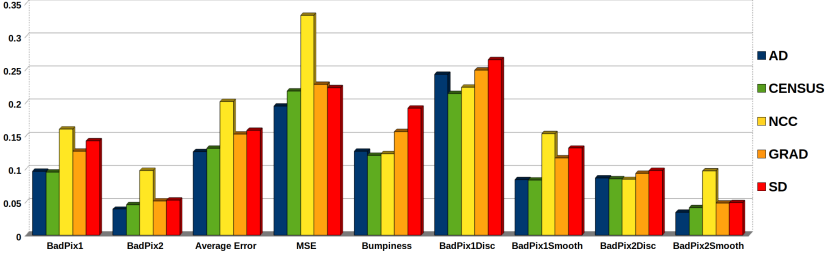


Figure 4.2. Synthetic images allow for a quantitative analysis of the effects of a similarity measures on the accuracy of the disparity estimation. *BadPixN* refers to pixels with an error higher than N , ($N = 1, 2$), *MSE* denotes the mean squared error, *Disc* and *Smooth* refer respectively to pixels belonging to depth discontinuities and to smooth regions in the scene. Modified image from Publication 2 in Section 6.2 [PKV18].

promising results in other fields, yet they underperform in the MI case, making them less suitable.

Even though choosing the optimal similarity measure or a combination of them reduces the noise in the cost volume, a second step is needed to guarantee higher accuracies. It consists in the regularization of the cost volume. A trade-off between quality and computational effort controls the choice of the algorithm: global and iterative methods could yield more accurate results, yet they are computationally expensive. For a fast yet precise estimation, the semi-global strategy [Hir07] described in Equation 4.2.2 is preferred.

$$\begin{aligned}
 C_{reg}(\mathbf{p}, d, \mathcal{L}_{ref}) = C(\mathbf{p}, d, \mathcal{L}_{ref}) + \sum_{\mathbf{r} \in \mathcal{D}} \min \{ & C(\mathbf{p} - \mathbf{r}, d, \mathcal{L}_{ref}), \\
 & C(\mathbf{p} - \mathbf{r}, d - 1, \mathcal{L}_{ref}) + P_1, \\
 & C(\mathbf{p} - \mathbf{r}, d + 1, \mathcal{L}_{ref}) + P_1, \\
 & \min_i C(\mathbf{p} - \mathbf{r}, i, \mathcal{L}_{ref}) + P_2 \}
 \end{aligned} \quad (4.2.2)$$

Where the regularized cost C_{reg} is obtained by recursively adding the minimum of the cost of the neighbouring pixels. The neighbouring pixels

4. Disparity Estimation using Plenoptic Cameras

are reached using \mathbf{r} that takes the directions \mathcal{D} , which account for the choice of the neighbouring pixels. Penalties P_1 and P_2 are used to enforce consistency between disparity labels of neighbouring pixels, with P_1 being used for pixels whose disparity only differs by one with respect to the reference pixel, and P_2 being used for the minimum of the disparities whose difference is higher, denoted as i in the equation. A common choice is $P_2 > P_1$, assigning higher penalties to larger label differences.

The final disparity map is extracted as the labels with the minimum cost across the volume, as described in Equation 4.2.3.

$$d_{final} = \arg \min_d C_{reg}(\mathbf{p}, d, \mathcal{L}_{ref}) \quad (4.2.3)$$

Neighbouring Microlens Selection

The advantage of MLA-based plenoptic cameras consists in having multiple correspondences for every point, ensuring more robustness in the matching process. However, finding and selecting the matching correspondences is a complex task, because the number of correspondences is depth dependent and not constant across the whole image [PW12]. Moreover, in the case of MPC different lens types has to be taken into account. Taking into account that recent plenoptic cameras have several thousands of microlenses, the process of selecting them for an accurate matching becomes challenging.

The aim of the proposed work on lens selection is to create an approach that allows a faster computation while preserving the information about the microlens structure and the accuracy of the disparity estimation. The idea takes inspiration from the concepts of training and calibration, having the most expensive computational steps done at the beginning in a trade to speed up runtime executions.

In this manuscript, the concept of virtual depth is used. The virtual depth is described in [PW12] as the number of adjacent microlenses along one epipolar axis that image the same point in the scene. For an object in the scene, the number of microlens it will be imaged onto is related to its virtual depth. This value requires the estimation of the disparity value. Therefore, selecting a microlenses combination before estimating disparity value is an ill-posed problem. Without disparity it is not possible to correctly predict in which microlenses the object appears, thus its

4.2. Disparity Estimation Algorithms

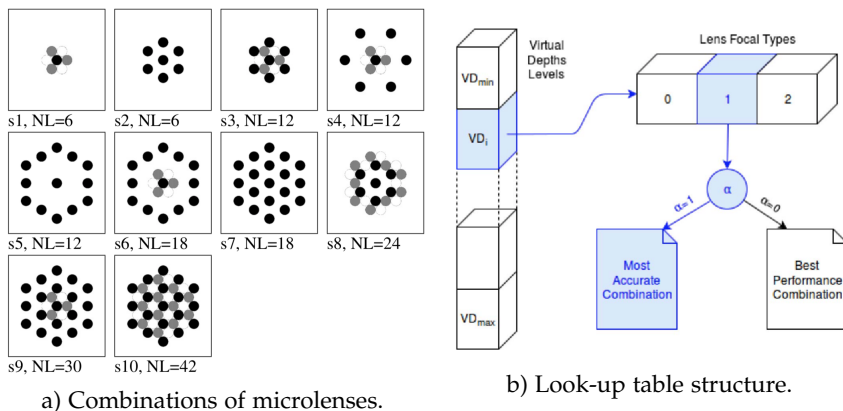


Figure 4.3. Different lens combinations have been evaluated and stored in a look up table to reduce the computational cost at runtime. Several combinations, along with the total number of lenses (NL) are shown in the image. The look up table shows a basic visualization of the algorithm at runtime, where combinations can be selected according to performance or accuracy. Image from Publication 1 in Section 6.1 [PK17].

virtual depth. Simultaneously this information is needed to choose the best combinations for the calculation of the disparity.

The chosen workaround divides the method into two steps. Since no information about the geometry can be assumed beforehand, the first step has to be independent from the position of the object in the scene. Only one assumption can be safely made, related to the lower bound, which is relatively straightforward. Since at least two correspondences are needed to estimate the disparity of a point by means of triangulation, it is assumed that the neighbouring lenses image the same object. A rough guess of the disparity is computed using a pre-defined lens combination based on the lower bound of the disparity range, thus on the neighbouring lenses or a subset of them. Then this value is used to select the optimal lens combination to refine the disparity to its final outcome. Different pre-defined lens combinations have been tested and the one that achieved higher accuracy has been selected.

In the second step the estimated disparity value is used to fetch the

4. Disparity Estimation using Plenoptic Cameras

best combination of neighbouring microlenses, which is used to refine the disparity estimation. At this step the disparity information is available, and assumptions on the disparity help reducing uncertainty: to ensure a correct estimation, it is assumed that the scene range falls within the correct optical range for the plenoptic camera. The upper bound does not depend on geometrical requirements, but on the microlens size and the spatial resolution. The higher virtual depth an object has, the more times it will be repeated. This implies being imaged with large angular and low spatial resolution. Objects with large virtual depth, for example $VD > 10$ for commercial implementations, are imaged at a very low spatial resolution and larger values often do not produce visually consistent results.

Based on these observations, a set of patterns was created, as visible in Figure 4.3. Each pattern took into account also the information about the lens type. To test the best configuration, sets of images perpendicular to the camera direction were created using the plenoptic simulator and a texture pattern. These images constitute a case study for a point at different depths in the scene, therefore giving us the possibility to analyze and measure the performances of those combinations of microlenses. The texture patterns is used to ensure a fair comparison and a robust disparity estimation, simulating the case of a textured region of the scene. The texture-less region remains an open challenge.

Experiments showed that our proposed approach is able to increase the accuracy of the disparity estimation while reducing the computational effort due to the matching, resulting in lower computational cost. Moreover, at run-time no estimation is needed since the combination is fetched from the look-up table. For more details about the lens combinations, the creation of the textured images and information about performances, please refer to Publication 1 in Section 6.1 [PK17].

4.2.2 Robust Estimation for Light Field Microscopy

While estimating disparity using a MPC deals with a huge amount of MIs, it may not be the case for different light field capturing devices. As explained in the previous chapter, the optical configuration regulates the sampling of the light field. Therefore, particular optical configurations need to have the methods adapted to work with different structures of information. In

4.2. Disparity Estimation Algorithms

this section, the estimation of the geometry of biological samples captured with a light field microscope is investigated.

By means of a self-made light field microscope from University of Valencia [SSL+18], light field of biological samples can be captured in real-time. Publication 5 in Section 6.5 [PSI+19] describes more in detail the work made to estimate depth maps from those images. Since these images have a different shape with respect to the MI from the previous section, a new dedicated method was developed.

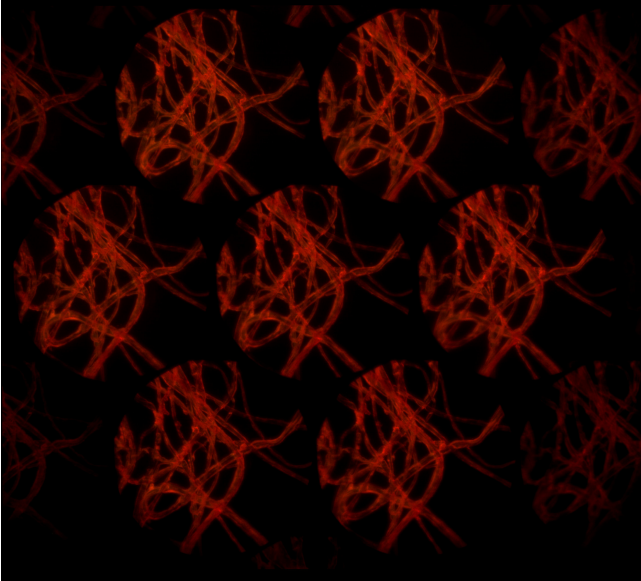


Figure 4.4. An example of an image captured with the light field microscope described in [SSL+18]. In this sample cotton fibers are imaged with a monocolored light. Image from Publication 5 in Section 6.5 [PSI+19].

A raw image acquired with the above mentioned microscope is shown in Figure 4.4. The microscope captures directly the SIS on the sensor, arranged on an hexagonal grid. Because of the circular aperture lens, the images also exhibit the same shape.

In this case, no transformation or preprocessing is needed to recover

4. Disparity Estimation using Plenoptic Cameras

the SI. On the other hand, the angular resolution is low and the parallax between each SI is larger. Moreover, estimating depth for biological sample constitutes a challenge because of the nature of the images. Their characteristics creates sub-optimal conditions for stereo matching: low light, absence of texture, uniform background and uniformly colored regions introduce challenges to establish robust correspondences. Figure 4.4 illustrates this. The cotton fibers are hit by a mono coloured laser and thus have a similar color with a dark textureless background.

In the proposed work a workflow was designed to deal with these challenges by combining different approaches for the depth estimation. As depicted in Figure 4.5, the pipeline includes many steps. It involves creating a cost volume, applying refinement step to reduce the noise and finally extracting the disparity label through a global optimization approach. This solution is chosen in this case because the importance of the accuracy outweighs the need for real-time calculations.

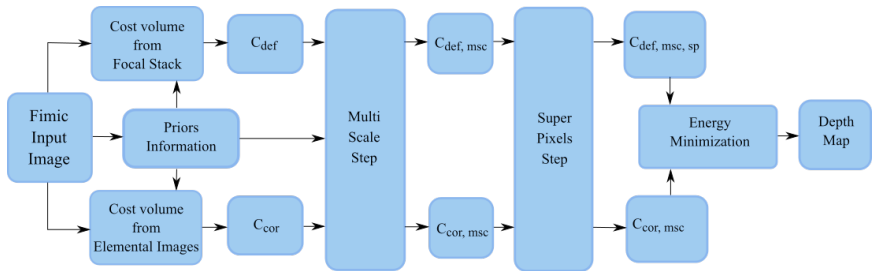


Figure 4.5. The diagram shows the workflow upon which the work was based. Several steps are used to guarantee a robust estimation for biological images. Modified image from Publication 5 in Section 6.5 [PSI+19].

Using this method, two different cost volumes are created. One is computed using dense stereo matching process and one comparing the central SI against a focal stack generated overlapping all the SIs. The creation of the focal stack is a characteristics of the images captured with the light field microscope and was investigated in previous works [SSL+18]. The priors information shown in diagram 4.5 relates to an additional mapping based on the epipolar lines, which is used to weight information in the stereo matching process and in the multi-scale step.

4.2. Disparity Estimation Algorithms

To ensure robustness against noise and match features of different sizes, two additional steps were introduced to address a wider range of images and features. The multi-scale cost volume is a modified version of the initial cost volume, where the computations are made on a set of images downscaled at different resolutions. Lower and coarser level works as a nudge towards the correct value, reducing the uncertainty in the finer estimation. The final value is a weighted average of the contributions from different levels, taking into account the prior information.

The same concept applies to superpixels. The difference consists in the formation of the superpixels, which are regions created by grouping pixels with similar characteristics, whereas downscaling is independent of the image content. Superpixels are created using the central SI and segment different parts of the scene. This is particularly important in the case of biological samples, where similarly colored foreground objects have to be distinguished from the uniform dark background. Similar values of depth are enforced within each superpixel through a penalty function, yet different values are allowed. The size of the superpixels is tuned depending on the application. Large superpixels lead to constant regions and small tend to have very little effects on the final estimation. The application of these additional steps reduce the noise in the cost volumes, allowing the extraction of a more accurate depth map.

The final extraction of the depth map is modelled as a multi-label optimization problem and solved through an energy minimization approach. It uses graph cuts where the energy function $\mathcal{E}(p) = \mathcal{E}_{data}(p) + \mathcal{E}_{smooth}(p)$ is adapted and the data term described in Equation 4.2.4:

$$\mathcal{E}_{data}(p) = (1 - \mathcal{M}_{ad}(p))\mathcal{C}_{def}(p) + \mathcal{M}_{ad}(p)\mathcal{C}_{cor}(p) + \beta\mathcal{P}_{gcp}(p) \quad (4.2.4)$$

Where the data term for a given pixel p retains the value of three contributions: the cost volume based on the defocus approach, indicated as \mathcal{C}_{def} , the cost volume based on correspondences matching, denoted as \mathcal{C}_{cor} and a penalty function based on ground control points \mathcal{P}_{gcp} scaled by a factor β . The term \mathcal{M}_{ad} refers to a pixel-based weights map, whose weights depend on the comparison between the two different depth maps computed using a winner-takes-all approach from the two cost volumes before optimization.

4. Disparity Estimation using Plenoptic Cameras

In the weights map \mathcal{M}_{ad} , larger weights correspond to pixels where the two cost volumes have similar minima, in order to rely stronger on the correspondence-based disparity, which tends to be more precise. The opposite is true for lower weights, where cost volumes take different shapes and the focus-based disparity tends to have lower errors.

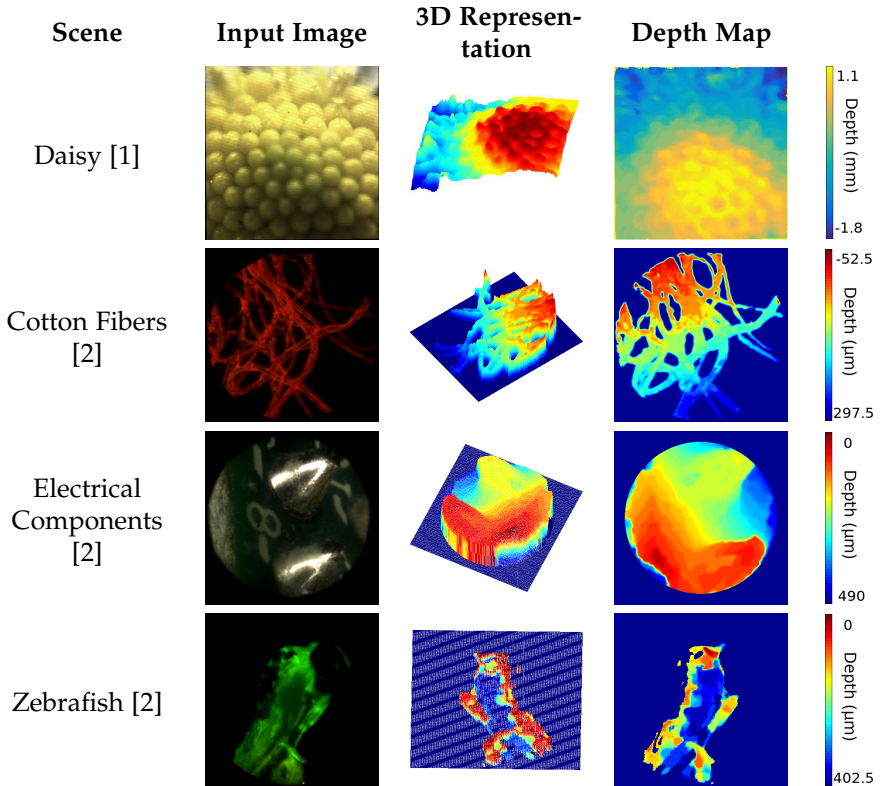


Figure 4.6. Images and respective results from different samples. It can be seen how the algorithm estimates depth with high accuracy at the micrometer scale, and a three-dimensional visualization is provided to highlight the complex structure of the scene. Samples marked with [1] were acquired in [MI15], while samples marked with [2] were captured in Valencia using the FiMic [SSL+18] as part of Publication 5 in Section 6.5 [PSI+19].

The algorithm has shown to be robust across a diverse range of input images, ranging from noisy low-resolution images acquired with a custom microscope setup in [MI15] to transparent biological sample and opaque electrical components imaged at higher resolution thanks to the light field microscope. Some results are shown in Figure 4.6.

For more details about the workflow, the implementation, the results and the comparison with the state of the art please refer to Publication 5 in Section 6.5 [PSI+19].

4.3 Experiments

Within the work presented in this thesis a framework to promote further research in the field has been created. Not all the experiments and projects carried out during this period have been published, yet they are basis for future works and of great help in reporting challenges found in the implementation. Thus additional experiments conducted during this period are reported here along with their preliminary results.

4.3.1 Disparity Estimation from the Focal Stack

In Section 3.3.4 the focal stack was introduced. This set of images can be used as input to estimate a disparity map. Using the same idea applied in Publication 2 (Section 6.2, [PKV18]) and Publication 5 (Section 6.5, [PSI+19]), a cost volume can be created by using a focus measure on the focal stack images to estimate the sharpness for each different focal plane. The rendering algorithm can select the step between planes in terms of patch size, which is proportional to the actual depth of an object in the scene. Since the rendering algorithm creates an arbitrary number of planes and it relies on tiling together patches extracted from MI, unsharp images often show strong artefacts, helping in the estimation of the correct focal plane at which each object in the scene is sharply imaged.

An experiment with a basic pipeline is described below and its results are shown in Figure 4.7. To investigate different approaches, in this experiment both the focal stack and the all-in-focus images were used, therefore using the disparity information. However, the disparity is not essential for

4. Disparity Estimation using Plenoptic Cameras

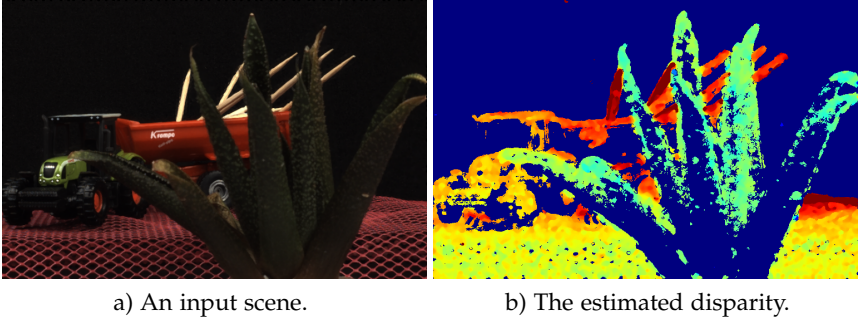


Figure 4.7. The image shows the output of a disparity estimation using the focal stack. Because of the noisy estimation, pixels with very low confidence were filtered out, leaving a sparser disparity. This accounts for a better visualization and understanding of the regions where the estimation fails. Images generated using the plenoptic toolbox [Pal20] described in Publication 2 in Section 6.2 [PKV18].

the estimation. After rendering a focal stack of $N = 20$ focal planes, the cost volume was created and smoothed using the bilateral filter. The cost of a pixel \mathbf{p} for a depth hypothesis at the focal plane d is calculated on the window \mathcal{W} centered on the pixel and the cost is computed using the focus measure \mathcal{F} , as described in Equation 4.3.1.

$$C_{def}(\mathbf{p}, d) = \sum_{q \in \mathcal{W}} \mathcal{F}(\mathbf{q}, d) \quad (4.3.1)$$

The focus measure \mathcal{F} can be used only on the focal stack or to compute the difference between the all-in-focus-image and each image of the focal stack. The best results were achieved using a sum of absolute difference between refocused images and the all-in-focus image. Using only the focal stack, comparable yet slightly less accurate results were obtained using difference of gaussians as the focus measure.

Finally, semi-global matching is applied to reduce the noise in the cost volume and the final extraction of the disparity labels is extended using a Taylor expansion to obtain continuous values and avoid large steps between discrete disparities.

Results show accurate reconstruction only around edges or textured regions and fails on textureless areas, as the black background.

Challenges, limitations and margins for improvement

The main bottleneck in this approach lies in the estimation of the sharpness of each slice of the image. If edges and corner are estimated with satisfying precision, textureless regions constitute a challenge and accounts for very noisy estimation. The problem is common to stereo matching approaches, therefore merging information from correspondences matching do not significantly improve the results. Moreover, at this preliminary stage, best results were achieved using the information from the all-in-focus rendered image, thus requiring prior estimation of the disparity. This can be solved by improving the focus measure based solely on the unfocused images, eliminating the need to render the all-in-focus image.

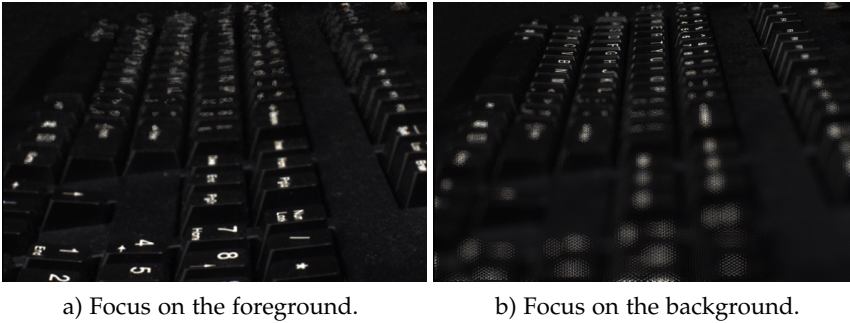


Figure 4.8. This scene is an extreme case, in which a large depth of field is required. When the foreground is sharp, the background exhibits blur similar to the gaussian one, while when the background is sharp, the foreground exhibits artefacts. This is due to the larger number of contributions from neighbouring lenses that overlap on this region. Images generated using the plenoptic toolbox [Pal20] described in Publication 2 in Section 6.2 [PKV18].

However, to improve the blur estimation, a further analysis of the behaviour of the blur should be beneficial. First, since the image is generated from the contribution of many neighbouring lenses, objects whose depth is very far from the focus plane show artefacts instead of gaussian blur. Moreover, because of the number of neighbouring lenses, blur generated at different depth show different forms, as visible in Figure 4.8. Although this is due to the implementation for the rendering of the focal stack and

4. Disparity Estimation using Plenoptic Cameras

could be corrected, it carries additional information that could be used to develop an ad-hoc focus measure to achieve higher accuracy. Secondly, the MLA in the case of MPC consists of three different lens types, that deliver further information. Exploiting the information about the defocus blur in different microlens types also gives a hint about the depth of an object in the scene. At the moment this is subtly incorporated in the creation of the focal stack, yet it could be analysed more in detail to improve the final estimation.

4.3.2 Disparity Estimation using Subaperture Images

As discussed in Chapter 3, the most widely used representation is based on the SIs. Since this constitutes a special case of multi view stereo, many approaches were proposed using this representation. Discussing all of them is outside the scope of this section, yet some experiments were made to compare the quality of the results using as input a set of SIs from a light field captured with both plenoptic camera models, SPC and MPC.

Since for the MPC case the disparity information is needed in order to render the SIs, the problem seems to be ill-posed. An evaluation of the disparity calculated on the SIs, however, implicitly carries out an evaluation of the goodness of the rendered images. Therefore it is interesting to evaluate the effort in rendering SIs from MPC and their similarity to the SI rendered from SPC. In these experiments two methods were used to evaluate those concepts. The first method corrects distortion on subaperture images, use them to create a cost volume which will be filtered to reduce its noise, and use an iterative global optimization step to improve the results [JPC+15]. The second is a learning based approach [SJY+18]. A neural network is trained on a synthetic dataset of light field images to estimate disparity using predefined sequences of images, in this case SIs with horizontal-only and vertical-only parallax and diagonally adjacent SIs where the ratio between vertical and horizontal parallax is constant.

Even though the methods are not part of the proposed work, it is interesting to see how the images generated with the proposed rendering algorithm, from scene acquired with MPC, compare with the images rendered using the state of the art light field toolbox [DPW13], from scene acquired with SPC. Results are encouraging, as visible in Figure 4.9.

4.3. Experiments

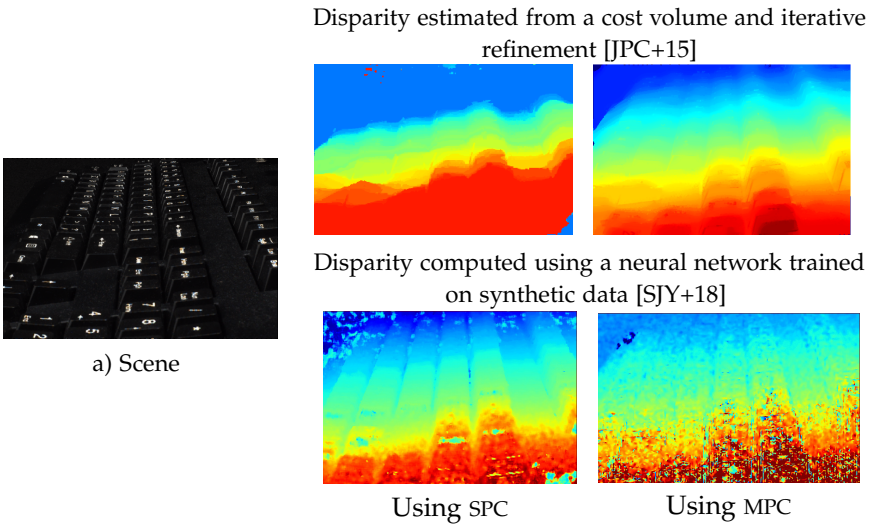


Figure 4.9. An example of a scene and the disparity maps calculated with both techniques. Parameters were tuned to aim for optimal results. The iterative refinement achieves a smooth shape at the cost of losing some features of the scene, while the EPINET approach retains the structure of the scene with high fidelity although delivering an overall slightly noisier image. Original image belongs to the dataset acquired in Publication 3 in Section 6.3 [APK+18].

In fact, the rendering algorithm is able to create a set of SIS which are very similar to the ones obtained using [DPW13] from SPC images. Tuning the parameters, it is possible to render images with similar disparity ranges and spatial and angular resolutions. The dataset acquired in Publication 3 in Section 6.3 [APK+18] is suitable for this experiment and SIS are created from those plenoptic images as the input for the two methods.

As expected, the methods return different results maintaining the similarities across cameras, confirming the quality of the rendered images. This particular image is an interesting case study, since it has diverse characteristics that give us an overview of strength and weaknesses of each algorithm: large depth of field, textureless black regions, areas with high contrast and small details. Presented results were, however, observed across different images. The iterative method achieves through the global

4. Disparity Estimation using Plenoptic Cameras

optimization steps a very smooth disparity map, at the cost of filtering out some details of the structure of the scene. Surprisingly enough, for this method, images rendered from MPC retain better performances. Since the network was trained on synthetic images, one expects noisier results when using the network on real images, and the outcome confirms it. Nevertheless, from a direct visual comparison is possible to notice how the structure details are actually better preserved in the second approach, at the cost of having noisier estimation. In this case, the MPC images constitute a less suitable input, exhibiting strong noise on the estimated disparity.

4.3.3 In the epipolar plane image domain

Another widely used approach for depth estimation is through the analysis of the EPIS. In this case, the procedure is quite different. As shown in Section 3.3.3, in the epipolar domain the depth is related with the slope of the lines in the two-dimensional image slices.



Figure 4.10. An horizontal epipolar plane image slice from Figure 3.5. Different slopes relates to different depths of the object. The image is scaled for visualization purposes.

In this domain it is convenient to represent the light field using the two-plane parameterization from the lumigraph [GGs+96], with one plane at the focal points of the views with coordinates (s, t) and the image plane with coordinates (x, y) .

$$EPI_{(y^*, t^*)} = L(x, y^*, s, t^*) \quad (4.3.2)$$

A two-dimensional slice can be then extracted by adding a constraint on two coordinates, as described in Equation 4.3.2 and visible in Figure 4.10.

Considering a three-dimensional point in the image, its depth Z is related to the slope of its line in the epipolar domain as described in Equation 4.3.3.

$$Z = f \frac{\Delta s}{\Delta x} = f \tan \theta \quad (4.3.3)$$

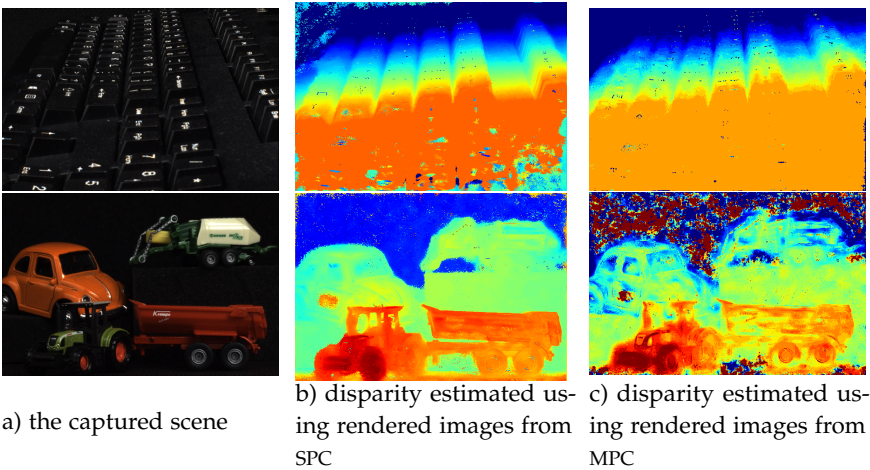


Figure 4.11. Disparity estimation from the EPI representation. The estimation is done using the spinning parallelogram operator approach described in [ZSL+16] on images acquired in Publication 3 in Section 6.3 [APK+18].

Where s and x relates to the two plane parameterization above described, f denotes the distance between the two planes and θ the angle or slope of a line. Therefore, the depth estimation task consists in calculating the angle θ . Different approaches have been proposed to guarantee robustness of the estimation: a local analysis using structure tensor is followed by a global optimization step in [WG12], a dedicated spinning parallelogram operator is devised in [ZSL+16] to combine depth estimation from different directions in the EPI.

As in the previous section, the aim of this experiment is to compare the rendered images and the performance of the same algorithm on images acquired with SPC and MPC. Results are visible in Figure 4.11 In this case the performance are similar: using the method described in [ZSL+16] the depth estimation is able to recover the overall structure of the image, yet it fails to correctly estimate fine structures and dark uniform areas. Since images captured with SPC has a larger angular resolution, these images are expected to be more robust in the estimation in the EPI domain and the results confirm the assumption.

4. Disparity Estimation using Plenoptic Cameras

4.3.4 Three-dimensional Reconstruction

Reconstructing a scene in three-dimensions is among the most interesting applications enabled by plenoptic image processing. Scenes acquired with a plenoptic camera are suitable for this purpose: by re-projecting the estimated disparity map through into its three-dimensional coordinates, one can obtain an accurate three-dimensional representation, which in case of an accurate metrical calibration resembles the real world geometry. This process is strongly affected from noise both in the estimation of the disparity map and the camera optics during calibration.

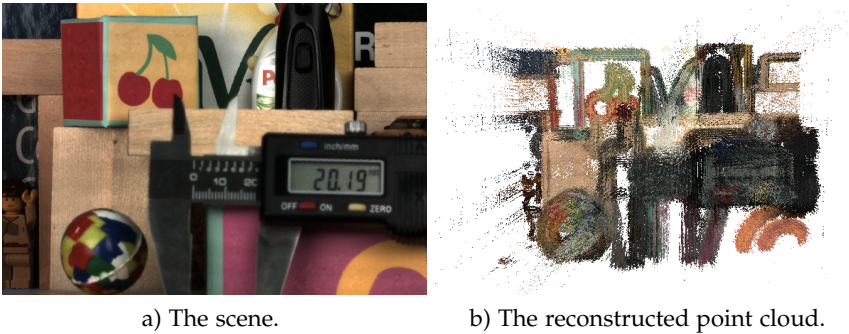
In the three-dimensional representation field, another interesting approach includes the fusion of multiple acquisition to reconstruct larger scenes. Starting from a three-dimensional representation of a single shot, it is possible to merge together different perspectives of the same scene to build an immersive 360 degrees three-dimensional model.

Metric Reconstruction

In order to metrically reconstruct the scene, a calibration is needed. The calibration consists in modelling the optical setup of a plenoptic camera. This can be done by decomposing the camera into two separate components, main lens and MLA, thus having two re-projection steps: one from the sensor to the virtual intermediate image and one from the virtual intermediate image to the real world coordinates.

Another way to do this is to use the H matrix that models directly the relationship between the disparity estimated on the sensor and the real world coordinates. This is the approach chosen in the Manuscript 1 in Section 7.1. By using the matrix, it is possible to transform the disparity map into its three-dimensional representation.

In this experiment, point clouds were chosen for a straightforward implementation and their absence of edge structures, which accounts for more flexibility and less computational effort. As visible in Figure 4.12, a point cloud can be reconstructed by reprojecting the disparity into three-dimensional points. In this case a thresholding step was introduced to remove noisy estimations. Since the projection relies on multiple steps, noise contributions are accumulated. A small error in the estimation



a) The scene.

b) The reconstructed point cloud.

Figure 4.12. The scene was captured using a Raytrix R42 (MPC) and reconstructed as a point cloud based on the calibration described in Manuscript 1 in Section 7.1.

translates into large errors in the three-dimensional space, therefore low confidence disparity estimations were filtered out, creating a sparser point cloud.

4.3.5 Full three-dimensional model

As mentioned above, being able to create a metrically correct three-dimensional representation enables the creation of a complete model where an object is imaged from different perspectives and fully reconstructed.

A project was set up to evaluate the requirements and the possibility to create such a model. The difference between using conventional cameras and plenoptic cameras for this task lies in the angular density needed to estimate the geometry of the scene. When using conventional cameras, large overlap is required to estimate disparity. Therefore hundreds of images (for example $N = 720$ images in [VAD+18]) are required for a full model, making the capturing process very slow.

Since plenoptic cameras enable a three-dimensional reconstruction from a single shot, no overlapping is needed. Therefore the requirements in terms of number of input images is significantly reduced. The setup of the experiment was designed to capture the scene from different perspectives while allowing maximum flexibility in the selection of the angles. Since

4. Disparity Estimation using Plenoptic Cameras

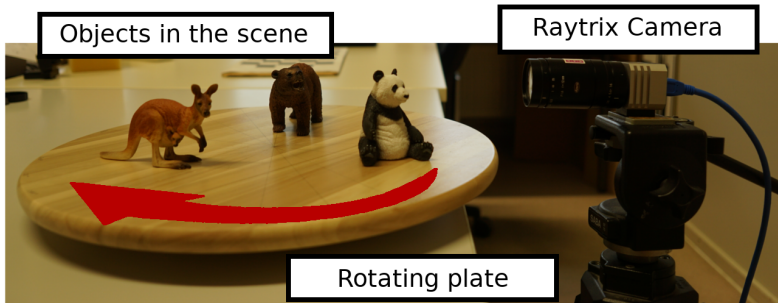


Figure 4.13. The setup of the experiment to capture the scene from different angles. The experiment was performed in the laboratory in University of Kiel.

the camera setup requires high precision, rotating the scene instead of the camera seemed the optimal solution. A rotating plate was used for this purpose, and the objects were placed in the scene and rotated, as shown in Figure 4.13. This allowed to capture the scene from different angles without moving the camera. The disadvantage of rotating the plate consists in the background, which does not remain constant as the scene rotates. To overcome this problem, in this experiment a green screen was used as background and a chroma-based segmentation was applied to retrieve the foreground object and remove the background.

The experiment was divided into two steps: first the disparity of each scene has to be estimated and reprojected in the world coordinates, then different scenes are merged together. For each scene a disparity map is calculated, from which a point cloud is created. Several scenes were acquired, creating a dataset with images captured every $\alpha = 15$ degrees. Since estimating the transformation between two different point clouds is a complex task, a constraint has been introduced. Shifting the reference system of the three-dimensional coordinates from the camera point to the middle of the rotating plate, the transformation between each camera pose consists only of a rotation, which is exactly the angle α the plate was rotated. In order to test the robustness and have a benchmark for the results, the same dataset can be reproduced with the plenoptic simulator. A 3D model of an elephant has been selected and rotated in the scene,



a) A snapshot of the point cloud reconstructed from a real scene.



b) A snapshot of the point cloud reconstructed from a synthetic scene.

Figure 4.14. Some results from the dataset. On the left, the point cloud created from a real image. Please notice that the background have been removed through color keying and outlier filtering. On the right, a point cloud obtained from a synthetically generated image using the simulator from Publication 4 in Section 6.4 [MPP+18]. The results on both sets look similar in terms of accuracy, with the estimation of synthetic images being slightly less noisy.

rendering the images with the Blender engine as described in Publication 4 in Section 6.4 [MPP+18]. The same dataset with images captured every $\alpha = 15$ degrees was created with an empty black background. Figure 4.14 shows a point cloud computed from both real and synthetic images.

Filtering along reprojection line

The principal challenge is to achieve a robust, clean and metrically consistent three-dimensional reconstruction. Although the result looks visually satisfying, the consistency between estimated measure in different axis is of extreme importance when trying to merge multiple point clouds whose relative transformation involves rotation. Therefore the noise in the estimation has to be reduced to the minimum.

To correctly reduce the noise, a dedicated filter was designed and

4. Disparity Estimation using Plenoptic Cameras

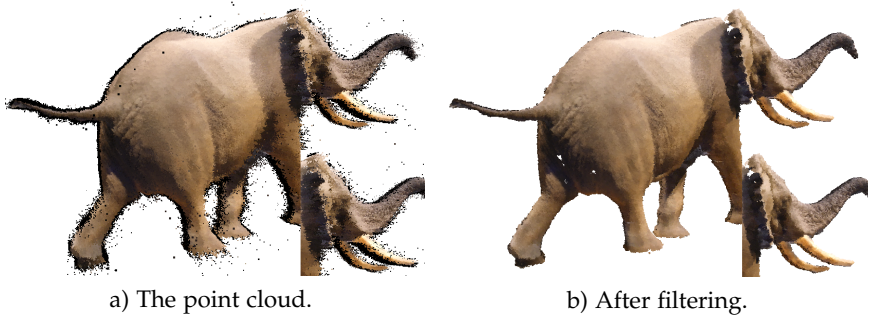


Figure 4.15. A point cloud estimated from a synthetic scene rendered using the simulator from Publication 4 in Section 6.4 [MPP+18]. The left image shows the point cloud after the reprojection from the disparity map. The noise is removed with a filter that selects the best value along the epipolar line.

implemented. Since these point clouds originate from the reprojection of thousand of micro disparity maps into real world coordinates, some assumptions on their position can be made. In the ideal case of an accurate reprojection, points belonging to objects visible in multiple MI will end up at the same three-dimensional location. The estimated disparities are subject to errors and noise, therefore will not end up overlapping. However, assuming a correct calibration, a noisy estimation of the disparity will mean that points which belong to the same object lie along the same reprojection line. Developing a denoising filter which works along the reprojection lines manages to significantly reduce the noise, creating a single clean filtered point cloud, as visible in Figure 4.15.

Challenges and limitations

Nevertheless, the fusion of multiple point clouds resulted in an increase of the overall noise and in misalignment issues, where a small distortion in one axis led to incorrect merging after several point clouds were added. Although the pairing of two point clouds is achieved using an iterative closest point approach or a similar method, the fusion of several point clouds to generate a full model could not reach satisfying results. The final implementation is still ongoing research and falls under future work.

Conclusions

This work is focused on various aspects of the processing of images captured with MLA-based plenoptic cameras. It covers several topics, from the requirements in the capturing process, the challenges in developing a unified representation for different sampling of the light field, the detection, extraction and clustering of features to achieve a metric calibration, and it maintains as core focus the disparity estimation, thoroughly analysed in its several aspects.

The main contribution of this work is twofold: on one side, it contributes to the plenoptic environment with a framework for working with plenoptic images and the possibility to convert these images into different representations, an optimal basis to build upon for future researches. The combination of available datasets, the toolbox to work with the images, the creation of synthetic images and the benchmarking based on ground truth data are a contribution to promote research in the plenoptic image processing field.

On the other side, the analysis addresses complex challenges in the disparity estimation of plenoptic images, investigating the potential and the limitations for different approaches. This work takes into account different MLA-based plenoptic camera models and their optical characteristics, studies the most suitable representations and includes an improved version of the rendering algorithm, which delivers high quality SIs from MPC images. Since MLA-based plenoptic cameras consists of several thousands of microlenses, the lens selection process has been optimized to find an efficient solution to select the best lens combination while improving the final accuracy. To improve the camera calibration, a dedicated algorithm has been designed, which is able to detect, cluster and extract corners, including information about its microlens type and the amount of blur. Benchmark-

5. Conclusions

ing was also targeted through the generation of synthetic images and the capture of the same scene with different plenoptic cameras, providing a solid ground for comparison and evaluation of diverse approaches. The estimation of the three-dimensional structure has been applied in the biological field, using images acquired with a light field microscope and developing an algorithm for robust depth estimation, which allows to reconstruct complex structures at a micrometer scale.

Because of their only recent hardware development, MLA-based plenoptic camera did not yet reach their full potential, and most applications show margins for improvement. Research in this direction keeps showing interesting results which hopefully will lead to a wider distribution of such technologies. The experiments described in Chapter 4 constitutes a basis for future works.

Publications

6.1 Publication 1

Optimizing the Lens Selection Process for Multi-Focus Plenoptic Cameras and Numerical Evaluation

Luca Palmieri and Reinhard Koch

Published in

Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 50-61). [PK17]

DOI: 10.1109/CVPRW.2017.223

Optimizing the Lens Selection Process for Multi-Focus Plenoptic Cameras and Numerical Evaluation

Luca Palmieri Reinhard Koch

Department of Computer Science, Kiel University, Kiel, Germany
{lpa, rk}@informatik.uni-kiel.de

Abstract

The last years have seen a quick rise of digital photography. Plenoptic cameras provide extended capabilities with respect to previous models. Multi-focus cameras enlarge the depth-of-field of the pictures using different focal lengths in the lens composing the array, but questions still arise on how to select and use these lenses. In this work a further insight on the lens selection was made, and a novel method was developed in order to choose the best available lens combination for the disparity estimation. We test different lens combination, ranking them based on the error and the number of different lenses used, creating a mapping function that relates the virtual depth with the combination that achieves the best result. The results are then organized in a look up table that can be tuned to trade off between performances and accuracy. This allows for fast and accurate lens selection. Moreover, new synthetic images with respective ground truth are provided, in order to confirm that this work performs better than the current state of the art in efficiency and accuracy of the results.

1. Introduction

The idea of plenoptic imaging was introduced by Lippmann in 1908 in [13], but only recent development have made it possible to actually build devices that are capable of capturing the so-called plenoptic function. In recent years the interest towards such devices is growing, and many different approaches are appearing. The micro-lens array based cameras, that are in fact equivalent to an array of cameras, as proved in [6], have mounted an array of micro lenses between the main lens and the sensor extending the capacity of the device to capture the light field in only one shot.

A specific subset of these cameras is characterized by the use of different micro lenses, particularly with different focal lengths, and they exploit this aspect obtaining for example a wider depth of field. Multi-focus plenoptic

cameras were recently introduced in 2012 by Georgiev and Lumsdaine in [5] and a detailed technical explanation of the multi-focus properties is present in [7], but they quickly attract interests in scientific research as shown by the publications addressing this specific cameras.

Different topics were tackled, like a robust automated calibration in [11] and [7], the lens selection and cost function description using semi global matching in [4], a faster feature matching approach in [3] and the whole pipeline until the 3D rendering in [9], but also in industrial application, as shown by the fast growth of companies exploiting the technology.

Such cameras can be used for entertainment as Lytro [14] is doing, or for inspection and modelling, like Raytrix [17] is doing, but also for photography related tasks, like the most recent approach brought by the Light company [12], consisting of a pocket-size camera that emulates the performance of a DSLR camera using multiple lenses with different focal lengths.

Our approach targets one of the mostly used model that accounts for three different types of lenses (with three different focal lengths) provided by Raytrix, but has the advantage of being quite flexible and could be applied to all devices that use different lens types and need a strategy to accurately and efficiently select each time which one to use.

1.1. Structure of the Paper

Section 2 reviews related works, to give the reader an overview of the state of the art techniques related to the topic; in Section 3 we show the specific case of the plenoptic cameras that we are using, and we make an assumption about the disparity estimation with a detailed motivation; in Section 4 we go through the first step of the proposed approach, that uses ground truth generated data to evaluate different combinations; in Section 5 we combine the results into a specialized structure that allows an easy and efficient execution of the lens selection algorithm; finally, Section 6 compares the results obtained against the previously known techniques both with synthetic and real data (acquired with a Raytrix camera) and Section 7 provides a conclusion and

6.1. Publication 1

some possible future developments. Appendix A is used to give the reader further insight on how synthetic and real data are generated.

2. Related Work

Many approaches have been proposed to address the challenge of creating accurate disparity maps from image captured with a plenoptic camera: we focus on the aspects that make this camera unique in his genre, the micro lenses array and more specifically the lenses and the characteristics of their usage.

To reconstruct the image from the light field as captured by the lens array, one needs to select multiple adjacent lenses and compute depth (or disparity) from them. The depth is needed to collect the correct light rays for the sharp real image.

The lens selection problem remains an open challenge of high importance, because it addresses the very nature of the cameras: the micro-lenses array that allows the camera to capture the light field in only one shot and controls the trade-off between lateral and spatial resolution, which adapts each camera for specific purposes like refocusing or estimating the disparity map and reconstructing the three dimensional geometry of the scene.

Previously proposed methods about lens selection always assume some geometrical information about the pixel, whose disparity or depth has to be computed, and are used to refine the estimation: they can be divided in two categories:

1. Using the geometrical information, limit the lenses range and check on every lens the amount of defocus blur and the minimum overlapping in order to understand if they could positively affect the estimation.
2. Divide the world into slices on the z-direction and assign at every slice a certain range of lenses.

The first approach was proposed by Fleischmann and Koch in [4], with an adaptive strategy that uses a first estimate of the disparity to select lenses, discarding the ones where the overlapping was without a certain threshold: their first estimation is efficient, but the adaptive strategy involves always some computational effort and does not reach the highest precision in the lens selection.

As an example of the second approach we pick the most recent paper on the topic from Ferreira and Goncalves [3], where they divided the space into four quartiles. When a point seemed to belong to a certain sector, they assigned the lenses range and a predetermined combination: the idea of dividing the space into slices is functional in terms of performances, since it does not involves further computations, but it lacks in accuracy, since they use only four different areas and three different combinations, and they discard many

lenses just because of the difference in the focal lengths, while those still may contain useful information.

The proposed approaches for the lens selection seems not to reach the optimal solution, lacking in terms of either accuracy or performances, mainly because of two issues that are common to this kind of data:

1. To predict which lenses should be used for a point or a lens, some geometric information about the position of that point should be known, and the accuracy of this information greatly affect the final result.
2. It's challenging to capture images with ground truth to evaluate different methods, due to the particularity of the cameras.

We address both problems and propose our solution in the following.

3. Initial Lens Selection

In this paper we deal with a very specific version of the multi-focus plenoptic cameras: the approach was developed using the Raytrix cameras with three different focal types and the lenses arranged in the hexagonal grid as shown in [4] and in Fig. 1.

The approach we propose is flexible, and it can be adapted to all multi-focus plenoptic cameras, where the lenses of different focal types need to be selected for the disparity mapping or any other application.

The method is mainly divided into two steps:

- First, we compute a virtual data set consisting of different known depth planes. Using this calibration data we test lens combinations that are optimized with respect to efficiency and accuracy.
- Next, we create a look up table structure where data is stored and can be used efficiently during runtime.

We will discuss this in details respectively in Section 4 and 5.

The mentioned calibration process is highly time-consuming, but it has to be performed only once and then the results will be stored in order to be used in every successive execution, in a similar way to a calibration process, allowing an efficient computation at runtime.

Before continuing, we briefly describe two concepts that are important for the rest of the paper.

Virtual Depth

The concept of *virtual depth*, introduced in [16], is defined as the number of different lenses in a row that image a point, so a point with virtual depth N , would be imaged by N different lenses belonging to the same line, i.e. N different horizontal viewpoints.

6. Publications

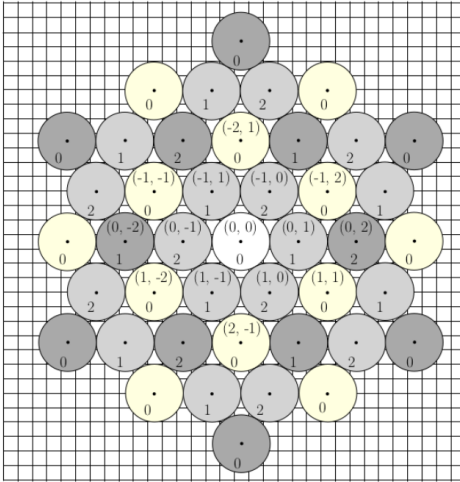


Figure 1: The lens grid of a multi-focus plenoptic camera: in this case a Raytrix camera with three different focal lengths is depicted. The numbers 0, 1, 2 indicates the focal type of each lens, and the coordinates (x, y) are relative to the central lens. Picture taken from [4]

Intuitively, the virtual depth is inversely proportional to the disparity: we can express that in mathematical terms in Eq. 1

$$VD = \frac{K}{d} \propto \frac{1}{d} \quad (1)$$

Where VD stands for the virtual depth, d for the disparity and K for a constant factor that is related to the metric calibration parameters of the lens array.

Disparity Estimation

Many techniques that computes the disparity are available. Based on the on the literature, we choose to use semi global block matching algorithm first introduced in [8] and used in [4], since it achieves better results as compared to the feature matching approach implemented in [3].

More sophisticated approaches exploiting the multi-view nature of these images will be inspected in future research.

3.1. Initial Virtual Depth Creation

Our method is based on the relation between a disparity and the combination of the lenses (that will lead to a refined version of the disparity), hence we need to compute a

first hypothesis on the position of the point in space, which does not have to be completely accurate. Since at this point we can trade accuracy for computational speed, we choose Fleischmann and Koch's [4] idea, making an initial guess using a small number of lenses and a block matching approach.

Nevertheless we reviewed different possibilities for this task: without changing the estimation method, we can select different combinations of lenses to reach a better results without loss of performances, as seen in Fig. 2:

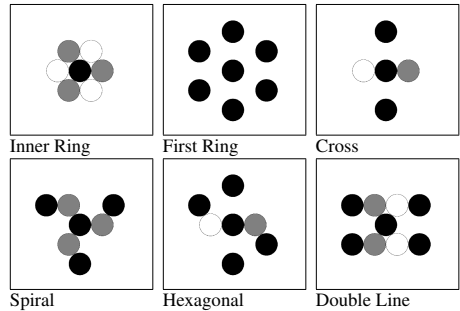


Figure 2: Different combinations of lenses that could be used for the initial virtual depth creation. They are chosen because of their structure around the central lens, and the gray colors represent their focal type: lenses with same color belong to same focal type.

The choices were made to tackle some particular characteristics of the lens grid: inner and first ring are most likely to be used because of the smaller baseline, being able to estimate disparity for both close and far objects.

Since the inner ring consists in lenses that have different focal lengths, they are not reliable, but necessary for close objects, whose virtual depth is small. The first ring, that contains only lenses with same focal type, thus with the same amount of defocus blur, should be more reliable.

We evaluated the different combinations on two datasets with respective ground truth, the first one also used in [4] to evaluate the results and the second one introduced to have more structure in the scene, in order to obtain both visual and numerical data to support our choice.

The visual feedback of both datasets is in line with the theoretical assumptions: the disparity computed with the inner ring is quite accurate for close points, but highly noisy for background points.

The opposite happens for the first ring, that addresses the far point in the correct way, but misses the correspondence for points with small virtual depth, as clearly visible in Fig. 3 and Fig. 4 where the centers of the micro images referring

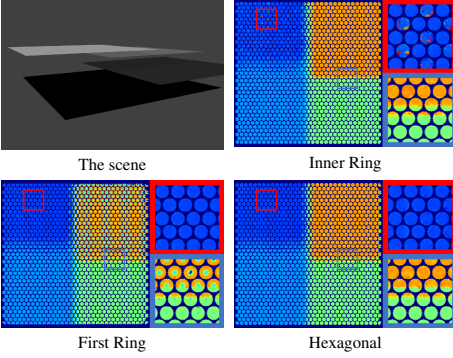


Figure 3: Evaluation on the first dataset consisting of four planes at different distances on the z-axis, to highlight errors at different disparities; only three of the combinations are shown here.

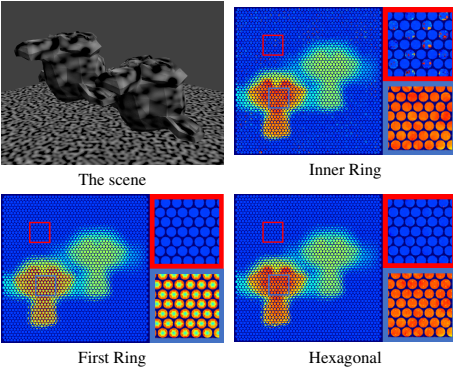


Figure 4: Evaluation on a more complex dataset where two objects show some structure, to see how each combination deals with border and objects close to each other.

to points close to the camera are not correctly computed, due to a large baseline.

We then see that the new proposed combinations use a mixed combination of both rings to address a larger depth of field, at a price of a lower accuracy.

Combinations with less lenses were tried with the scope

The disparity maps are shown in colored version for an easier visualization: red color means high disparity, and blue indicates lower disparity values. The two datasets are synthetically generated as explained in Appendix A.2.

of reducing the computational cost and increase the performances, but as shown the estimation with only four lenses is noisy and would affect negatively the final estimation.

Combinations with the same number of lenses show a better quality, particularly the best results are achieved for the *hexagonal* strategy, that uses six lenses as depicted in Fig. 2 (Hexagonal): two lenses with different focal lengths (white and gray) and four lenses of the same type around the central one. The *double line* combination achieve even more accurate outcomes, but at a price of a higher number of lenses.

Combination	Four Planes		Objects		Average		Lenses
	Avg.	Std.	Avg.	Std.	Avg.	Std.	
Inner Ring	0.305	0.568	0.427	0.851	0.366	0.710	6
First Ring*	0.359	0.664	0.319	0.666	0.339	0.665	6
Cross	0.267	0.438	0.319	0.671	0.293	0.555	4
Spiral	0.254	0.402	0.284	0.562	0.269	0.481	6
Hexagonal	0.244	0.387	0.282	0.392	0.263	0.390	6
Double Line	0.237	0.364	0.273	0.570	0.255	0.467	8

Table 1: Combinations and respective errors. Values are expressed in pixels and the disparity range is [0.5, 12.5]. We report here more combinations, to show how different approaches would deal with the problem.

* = combination used in [4]

The errors reported in Table 1 are computed using a simple absolute difference function, taking into considerations only valid pixels: for every micro-image only the pixels contained in a circle with a diameter slightly smaller than the image side are taken into account, with the exact value of the diameter set during the calibration process to avoid vignetting errors.

We print both average and standard variation to give an idea of how the error is distributed: a smaller variance would translate into a more robust outcome, that would be preferable in our case, since large error could lead to a wrong lens selection and thus to a wrong final estimation.

Analyzing both visual and numerical results, we select the *hexagonal* combination, that seems to solve at the best the trade-off between performances and accuracy.

We use this selection in the rest of the paper, so that every time we refer to the initial disparity guess, we mean the value obtained as explained above with the *hexagonal* combination.

4. Optimizing the Lens Selection

Based on the previously discussed assumption, we are now ready to proceed to the second step: having an initial guess of the point position allows us to create a direct mapping from this information to the best possible relative com-

6. Publications

bination of neighbouring lenses to fully exploit the characteristics of the micro lens array.

The novelty of this approach consists in the creation of a ground truth set of data by simulating the array of cameras: we set the position of our camera and we generate the images of a plane at a certain distance z from the camera, as if it was capture from a multi-focus plenoptic camera, using the right focal length for each micro-lens. To see the details about the creation of these synthetic planes, we refer to Appendix A.2.

Moving this synthetic plane from close to far with respect to the camera allows us to create a test set for all different positions that a point could assume in space and its corresponding ground truth, since we know the exact distance between plane and the virtual camera position.

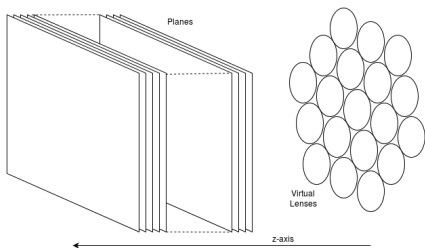


Figure 5: Representation of the planes and the lenses.

At this point we also want to stress that our specific cameras use lenses with three different focal lengths, so we need to evaluate the error and the combinations per each lens type, i.e. taking into account the amount of defocus blur of each particular lens.

An important parameter that has to be tuned is the distance between each plane: evaluating our combinations on planes that are too close to each other will result in erroneous selection, since our initial disparity guess cannot be particularly accurate, and using planes too far away will use the same combination for points that could benefit from a different one.

Our choice is to use the virtual depth measurement, as explained above, which has the advantage of being scale-independent, and extends the flexibility of our approach, leaving space also for particular application when the range of the scene has specific constraints and we are not interested in the whole depth of field of the camera.

We have chosen to use textured planes because of they spatial structure, being able to divide the world space in slices, and since the disparity estimation is not based on the structure, but just on the intensity of the pixel, textured planes fit our requirement for this task.

4.1. Comparisons among Combinations

The planes dataset was used to run all the disparity estimation, and every different combination was evaluated using the same error function explained above, absolute difference with respect to the ground truth, discarding the border of the lens.

We evaluated several combinations through the whole range of the scene, and we report here the ones that gave the most significant outcomes.

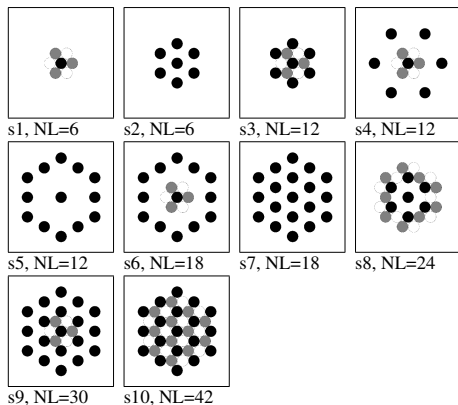


Figure 6: Different combinations: central lens is always shown in black, and for the other lenses every color represent a different focal type. The number of lenses (NL) is also reported.

As expected, the combinations that use adjacent lenses report better results on the close range (from 2 to 4 virtual depths) but show a large error when the distance from the camera increase, resulting useless in those cases; the opposite happen for the combinations that uses lenses with a larger baseline, as they need at least between 4 or 5 virtual depths to start working properly.

Once we calculated the results, we can sort them based on the average of the error in the disparity image and pick the best combination for every value of virtual depth. One can see that different combinations obtain similar results in the central area, where many different combinations are possible and many correspondences are available, thus increasing the difficulty of a correct choice.

Our idea is to develop two possible combinations for each slice: one which would give the most accurate result and one which would use the lowest possible number of lenses given a certain error threshold, in order to boost performances keeping a low error.

6.1. Publication 1

VD	Best Combinations		Perform.		Accuracy	
	LT	CBs	CB	NL	CB	NL
2	0	s1	s1	6	s1	6
	1	s1	s1	6	s1	6
	2	s1	s1	6	s1	6
3	0	s1	s1	6	s1	6
	1	s1	s1	6	s1	6
	2	s1	s1	6	s1	6
4	0	s2, s3	s2	6	s3	12
	1	s2, s3, s8	s2	6	s3	12
	2	s2, s3	s2	6	s3	12
5	0	s2	s2	6	s2	6
	1	s2, s3	s2	6	s3	12
	2	s2, s3, s8	s2	6	s3	12
6	0	s5, s2, s7	s2	6	s5	12
	1	s4,s6,s2,s3	s2	6	s3	12
	2	s10, s3, s8, s9	s3	12	s8	24
7	0	s5, s2, s7	s2	6	s5	12
	1	s2, s3, s7	s2	6	s3	12
	2	s2, s3, s8	s2	6	s8	24
8	0	s5, s2, s7	s2	6	s5	12
	1	s5, s2, s7, s9	s2	6	s9	30
	2	s2, s8	s2	6	s8	24
9	0	s5, s7	s2	6	s7	18
	1	s10, s2, s7	s2	6	s7	30
	2	s10, s2, s3, s8	s2	6	s8	24
10	0	s2, s7	s2	6	s7	18
	1	s5, s6, s3, s9	s3	12	s9	30
	2	s2, s3, s8, s9	s2	6	s7	18
11	0	s2, s7	s2	6	s7	18
	1	s3, s10	s3	12	s3	12
	2	s2, s3, s7, s8	s2	6	s8	24
12	0	s6, s9	s6	18	s9	30
	1	s3, s9	s3	12	s9	30
	2	s2, s3, s8, s9	s2	6	s9	30
13	0	s6, s9	s6	18	s9	30
	1	s3	s3	12	s3	12
	2	s2, s3, s8	s2	6	s3	12
14	0	s6, s2, s7, s9	s2	6	s9	30
	1	s3	s3	12	s3	12
	2	s2, s3, s8	s2	6	s3	12

Table 2: Outcomes of the lens selection step: the best combinations that satisfy Eq. (2) are grouped in the third column; choices for best performance and accuracy are shown in the columns on the right side.

Legend	
VD	Virtual Depth
LT	Lens Type
CB	Combination
NL	Number of Lenses

Table 3: Legend for Table 2

To retrieve such combinations, the outcomes are ranked for accuracy, then all combinations that satisfies Eq. (2) are grouped and sorted this time based on performance, using

the total number of lenses used for the estimation.

$$\mu_{f_j,i} < 1.5\mu_{min,i} \quad (2)$$

Where $\mu_{f_j,i}$ is the mean of the error for the j-th combination and for virtual depth i and $\mu_{min,i}$ is the minimum error for virtual depth i achieved by any of the combinations.

The results reported in Table 2 show the difference between the lens types and their need of different combinations in order to reach the highest accuracy.

The reported best combinations have similar outputs, so other choices are also possible, based on particular scenes or different parameters when acquiring or generating the scene, or also introducing different constraints on errors or maximum number of lenses. However, in our case we found these combinations to be the best.

The idea we want to highlight here is the fact that every lens, based on his focal type, could benefit a different patch of lenses for the disparity estimation: moreover, as the virtual depth increase, combinations that exploit lenses with higher baseline seems to achieve more accurate estimations.

Finally, it can be noticed that the different amount of blur highly affect the estimation of disparity images: as also found in [3], most of the lenses that will actually be used belong to the same lens type of the central lens.

5. Storing the Lens Selection

Once we gathered the results of our simulation, we need to store them in a way that allows us to use it in every next execution of the software: many structures could be used for this purpose, but our choice has fallen on a structure that resemble the characteristics of a look up table, that allows us to retrieve the best possible combination for each lens and lens type.

5.1. Choice of the Structure

The look up table structure seems to be the best solution is this case for different reasons, namely:

- **Efficiency:** once we have computed the initial guess, the computational effort needed to retrieve the relative position of the lenses of a particular combination from a virtual depth value is only a fetching instruction
- **Flexibility:** using virtual depth we have a measure that is scale-independent and the same structure can be used for all subsequent acquisitions.
- **Different choices available:** depending on the specific task, the user may want a different outcome; if quality of the results is the primary concern, the best combination of available lenses is selected, but if the operation to be performed has more speed constraint

6. Publications

and need a quicker execution, a parameter controlling the choice, changes the selection towards the quickest combination (i.e. the combination that uses lowest number of lenses and still achieve an error lower than a certain threshold)

5.2. Creation of the Structure

Based on these ideas, we created a structure that can manage the trade-off between accuracy and performance, storing both combination at the same time, and giving as output only one of them when needed

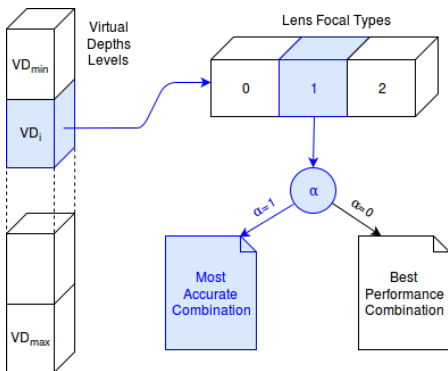


Figure 7: Look up table structure, with an example of a point with an initial guess of VD_i , belonging to lens with focal type 1, with parameter α optimized for accuracy.

The table has D levels, where D is the number of *slices* in which we divide the space (in our case $D = 13$) that have LT entries each, where LT is the number of lens types present in the camera (in our case $LT = 3$): those entries contain the relative combination of the lenses to be used for the estimation, and a parameter α is used to select which combination (most accurate or best performance) will be used, as depicted in Fig. 7.

If this approach had to be adapted to a different type of plenoptic camera, a simple change of the parameters would be enough to use the same structure in any other case.

5.3. Runtime Execution

Assuming we have calibrated our camera and created the look up table, we also create our internal mapping from disparity to virtual depth values, knowing that the virtual depth (VD) is proportional to the inverse of the disparity (d) as shown in Eq. (1).

The execution at runtime is controlled by a simple lookup: we feed as input three values, namely the virtual

depth (VD), the lens type (lt) and the parameter controlling the trade-off between accuracy and performances.

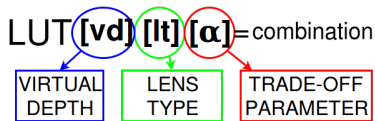


Figure 8: Graphic representation of the function that controls the look up table.

The output of such function are the relative positions of the lenses that should be used to achieve the best results during disparity estimation.

6. Results

We show some results to evaluate the proposed approach with respect to the state of the art technique; we use both synthetic images with generated ground truth to get some numerical value as an objective evaluation and real images without ground truth as a visual subjective evaluation.

The achieved results show an improvement of the accuracy of the estimation in some problematic areas, namely border of irregularly shaped objects, smooth textureless areas and fine structures, that are challenging for the task of disparity estimation.

We moreover stress the fact that this improvement is achieved without any additional computational efforts: while the calibration step and the creation of look up table is costly and time-consuming (but it has to be done only once), at runtime execution the algorithm does not need to compute any calculation and is able to choose the lenses to be used with only a simple fetching instruction in the look up table, resulting in an efficient approach.

Generally the selected combination use a high number of comparisons because we are looking for an accurate disparity map, but due to the flexibility of the proposed approach we can also change the orientation towards a more performant approach changing the threshold value that creates the different combination in the look up table, allowing the targeting of different applications.

6.1. Real scenes

The reported images are excerpts of the full scenes from Raytrix [17] and focus on certain details that we want to focus on; we compare them with the previous method implemented in [4] and with the results extracted from the Raytrix RxLive software [18]: the parameters used to obtain such disparity map per lens are tuned towards a more dense result, that is not necessarily the best result, but it's important in our comparisons to highlight the areas where

6.1. Publication 1

Scene	Computed with [4]		Ours	
	Avg.	Std.	Avg.	Std.
Four Planes				
Lens type 0	0.29	0.56	0.27	0.48
Lens type 1	0.26	0.44	0.23	0.33
Lens type 2	0.26	0.49	0.23	0.35
Platonic				
Lens type 0	0.47	0.41	0.46	0.41
Lens type 1	0.44	0.40	0.42	0.40
Lens type 2	0.41	0.40	0.40	0.40
Tomb				
Lens type 0	0.28	0.27	0.24	0.25
Lens type 1	0.27	0.25	0.23	0.23
Lens type 2	0.28	0.23	0.22	0.19

Table 4: Mean Error and Standard Deviation for the three synthetic scenes. Values are expressed in pixels.

the estimation is most challenging, without the successive filling algorithm.

The first scene, shown in Fig. 9, is the widely used Watch scene and gives us a good example of how a different selection of the lenses can affect the final estimation of challenging areas, like the textureless background surface and the textured plane, exhibiting a high level of noise that is reduced in our implementation, obtaining a more robust and smooth estimation.

Second scene, Fig. 10, consists in a more challenging outdoor scene, with a zoomed area relative to the left hand of the girl in the front, where the improvement with respect to the previous approach is quite small, but is possible to notice how the small details that cannot be reconstructed with enough reliability from the RxLive software are computed with a high accuracy and detail.

Fig. 11 finally highlights the smoothness and robustness of the estimation for detailed and highly textured areas, where the precision is raised and the noise in the estimation is almost removed.

6.2. Synthetic Scenes

The synthetic images consist in a fundamental step for this kind of disparity per lens images: up to our knowledge, no other methods were proposed to produce a numerical output to measure accuracy of the final estimation.

The images are more trivial and do not yet reach the complexity of a real scene: this is a task that we are currently addressing for the future to extend our qualitative results, but are still very helpful for evaluation purposes at the moment.

The scenes are part of the dataset developed in [10] and available at the 4D Light field Benchmark website [15], but

due to our settings, they have different point of view and disparity ranges.

The difference between the two estimations are not large and can appear unclear at a first glance, but as is visible from Table 4 our approach reduces slightly the error and obtains a lower standard deviation, meaning the estimation is more robust.

7. Conclusion

In this paper we tackled an issue common to multi-focus plenoptic cameras that represent still an open problem: our contribution is not only important in terms of positive quality of the results, but also in terms of the characteristics of the approach.

It takes from the idea of training the camera to achieve higher accuracy in the final outcomes, trading a computationally expensive lens selection phase to be done once allowing a multiple times more efficient runtime execution, that up to our knowledge was not proposed yet in terms of lenses estimation.

This approach works in this specific environment due to our assumption relative to an initial disparity guess and due to the nature of the problem: the lens selection is performed on the estimated distance in the z-axis of the point, without relying on texture or color information, therefore the training step can be done on planes while the execution process will most likely be done on different shapes without changing the outcomes.

Since in the last years these kind of camera are developing a high potential for different number of applications, as pointed out by the recent introduction of the L16 Light camera that exploit many different lenses with three different focal lengths for enhanced photography, hence the lens selection process is worth a further insight to exploit the full potential of the cameras.

Secondly, we were able to provide some light field images for a numerical evaluation, a missing element in the disparity per lens estimation field, where, as shown in [3] and [4], apart from a really basic scene, only a visual evaluation was possible.

We start to introduce new images and we look forward to building a small dataset to allow different estimation techniques to be compared.

Next step would be to focus on the comparison between those lenses, trying to evaluate which would be the optimal similarity measure or methodology to be used for that purpose.

6. Publications

6.1. Real Scenes

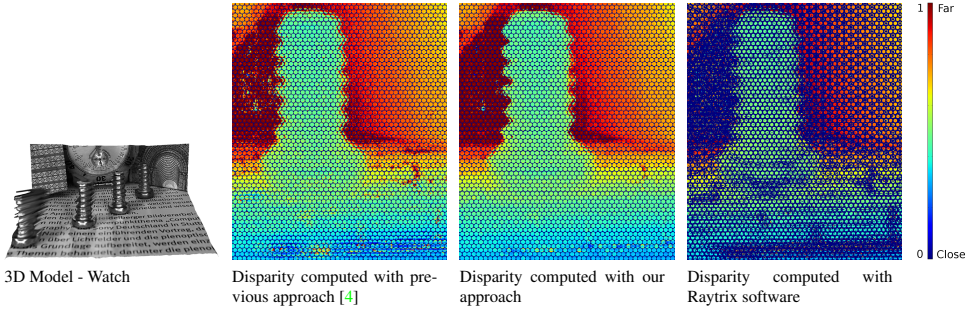


Figure 9: Part of the Watch scene.

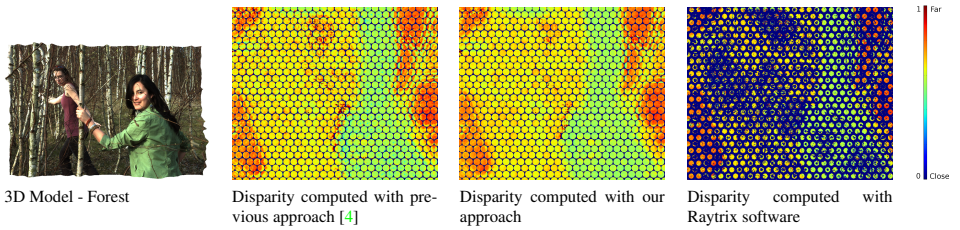


Figure 10: Part of the Forest scene.

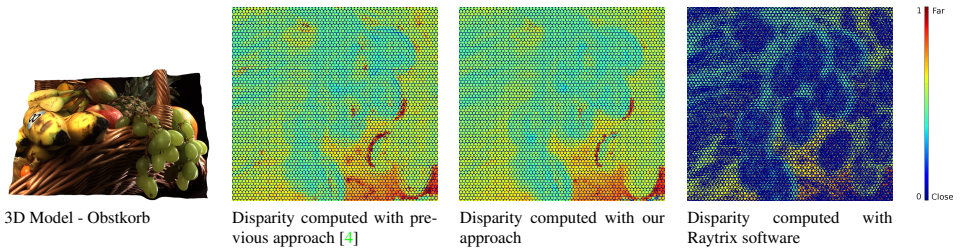


Figure 11: Part of the Obstkorb scene.

Scenes are provided by Raytrix [17].

6.2. Synthetic Scenes

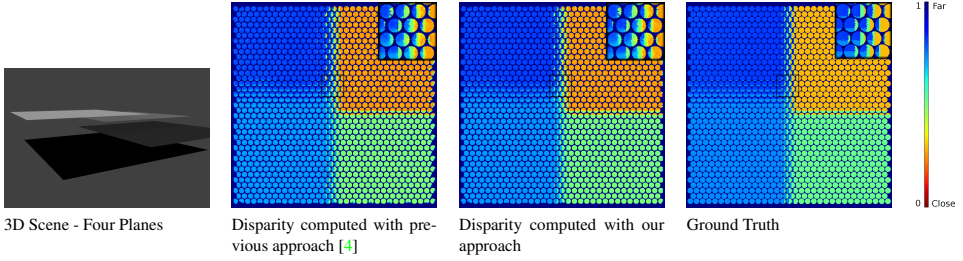


Figure 12: The Four Planes scene.

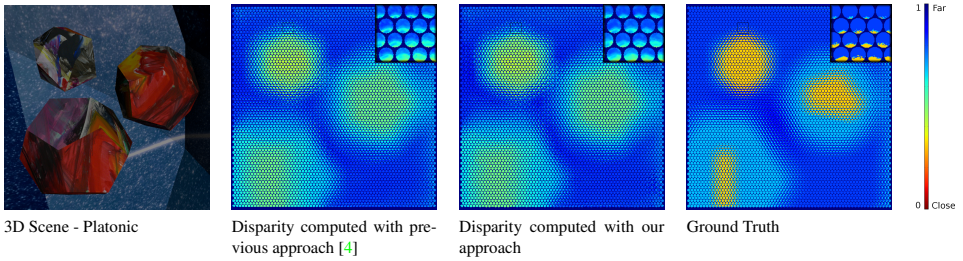


Figure 13: Part of the Platonic Scene.

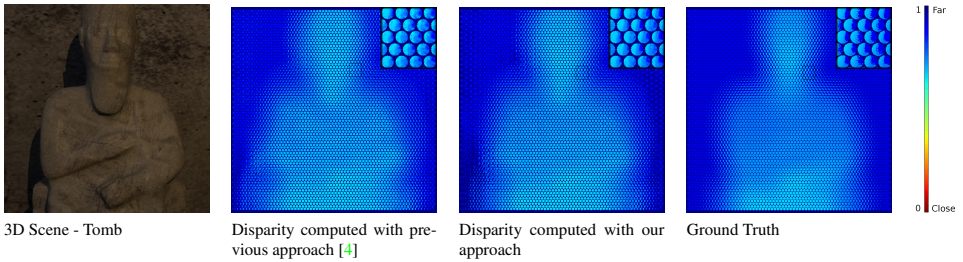


Figure 14: Part of the Tomb Scene.

The first scene (Four Planes) was produced by O.Fleischmann in [4].

The last two scenes (Platonic and Tomb) are part of the 4D Light Field Dataset [15] and the 3D models were used to generate the images with the respective ground truth.

6. Publications

A. Data Generation

Here we report more details about the generation of real and synthetic scenes with ground truth data.

A.1. Real Scenes

The real scenes were captured using Raytrix cameras. The 3D models and the disparity maps were extracted using the RxLive4.0 software provided by Raytrix [17].

A.2. Generating Synthetic Data with Ground Truth

The generation of the ground truth is a more complex process: since it's not yet publicly available any dataset with ground truth of disparity map per lens, evaluations is a challenging issue.

We provide ground truth images that simulates the image acquisition, even though they are not completely the same: the imaging process of Raytrix camera consists in projecting the real world scene through the main lens to an intermediate image in a virtual space, that stretches the relative depths and makes it easier to estimate larger disparities values. The particularity of their technique is that the intermediate image is virtually projected behind the micro lenses array, in a counter intuitive way not possible to reproduce with other cameras.

Our simulated images use the same idea, recreating the situation where the a virtual micro lenses array see the intermediate image, but this image is in fact in front of the array of micro lenses. The virtual depth of an image is thus inverted, meaning that an object with a large virtual depth would be close to the camera in a real image, and far away in a synthetic generated image, but since we focus on the mapping from virtual depth to lenses combination, these images fit perfectly our needs.

This idea was already exploited in [4] for a numerical evaluation, we extended this to scenes that consist in real benchmark for light field disparity estimation by recreating them with our synthetic generation pipeline.

Acknowledgment

The work in this paper was funded from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

Moreover the authors would like to thank O. Johannsen (University of Konstanz) for sharing some of the scenes of the benchmark for light field evaluation, A. Petersen (former Raytrix GmbH) for sharing scenes captured with Raytrix cameras and O. Fleischmann (former University of Kiel) for the contribution in the generation of the synthetic scenes.

References

- [1] M. Damghanian, R. Olsson, M. Sjostrom, A. Erdmann, and C. Perwass. Spatial resolution in a multi-focus plenoptic camera. In *IEEE International Conference on Image Processing (ICIP)*, 2014.
- [2] J. R. Bergen E. H. Adelson. *The Plenoptic Function and the Elements of Early Vision*. Cambridge, 1991.
- [3] R. Ferreira and N. Goncalves. Fast and accurate micro lenses depth maps for multi-focus light field cameras. In *German Conference on Pattern Recognition (GCPR)*, 2016. 1, 2, 3, 6, 8
- [4] Oliver Fleischmann and Reinhard Koch. *Lens-Based Depth Estimation for Multi-focus Plenoptic Cameras*, pages 410–420. Springer International Publishing, 2014. 1, 2, 3, 4, 7, 8, 9, 10, 11
- [5] T. Georgiev and A. Lumsdaine. The multi-focus plenoptic camera. In *SPIE Electronic Imaging*, January 2012. 1
- [6] T. Georgiev, A. Lumsdaine, and S. Goma. Plenoptic principal planes. In *Imaging Systems and Applications (IS)*, *OSA Topical Meeting*, July 2011. 1
- [7] C. Heinze, S. Spyropoulos, S. Hussmann, and C. Perwaß. Automated robust metric calibration algorithm for multifocus plenoptic cameras. *IEEE Transactions on Instrumentation and Measurement*, 65 (5), May 2016. 1
- [8] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, February 2008. 3
- [9] M. Hog, N. Sabater, B. Vandame, and V. Drazic. An image rendering pipeline for focused plenoptic cameras. *Hal*, 2016. 1
- [10] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision (ACCV)*, 2016. 8
- [11] O. Johannsen, C. Heinze, B. Goldluecke, and C. Perwass. On the calibration of focused plenoptic cameras. In *GCPR Workshop on Imaging New Modalities*, 2013. 1
- [12] Light. <https://www.light.co/>. 1
- [13] G. Lippmann. Epreuves réversibles. photographies intégrales. *Académie des sciences*, pages pp. 446–451, 1908. 1

6.1. Publication 1

- [14] Lytro. <https://www.lytro.com>. 1
- [15] University of Konstanz and Heidelberg Collaboratory for Image Processing. <http://hci-lightfield.iwr.uni-heidelberg.de/>, May 2017. 8, 10
- [16] C. Perwaß and L. Wietzke. Single lens 3d-camera with extended depth-of-field. In *Proceedings of SPIE - The International Society for Optical Engineering*. Addison-Wesley, February 2012. 2
- [17] Raytrix. <https://www.raytrix.de>. 1, 7, 9, 11
- [18] RxLive Software. <https://www.raytrix.de/downloads/>. 7

6. Publications

6.2 Publication 2

The Plenoptic 2.0 Toolbox: Benchmarking of Depth Estimation Methods for MLA-Based Focused Plenoptic Cameras

Luca Palmieri, Ron Op Het Veld and Reinhard Koch

Published in

2018 25th IEEE International Conference on Image Processing (ICIP) [PKV18]

DOI: 10.1109/ICIP.2018.8451073

THE PLENOPTIC 2.0 TOOLBOX: BENCHMARKING OF DEPTH ESTIMATION METHODS FOR MLA-BASED FOCUSED PLENOPTIC CAMERAS

Luca Palmieri, Reinhard Koch*

Department of Computer Science
Kiel University, Germany
{lpa, rk}@informatik.uni-kiel.de

Ron Op Het Veld[†]

Fraunhofer IIS
Erlangen, Germany
ron.ophetveld@iis.fraunhofer.de

ABSTRACT

MLA-based focused plenoptic cameras, also called type 2.0 cameras, have advantages over type 1.0 plenoptic cameras, because of their better inherent spatial image resolution and their compromise between depth of focus and angular resolution. However, they are more difficult to process since they require a depth estimation first to compute the all-in-focus image from the raw MLA image data. Current toolboxes for plenoptic cameras only support the type 1.0 cameras (like Lytro) and cannot handle type 2.0 cameras (like Raytrix). In addition, there is a lack of ground truth data and high quality benchmarking data for focussed plenoptic cameras. This contribution will discuss the requirements for processing type 2.0 images and will supply the reader with an open-source toolbox for comparing depth estimation methods. Different depth-estimation methods for MLA-based imaging will be available and an easy extension for other processing algorithms like compression will be included. In addition, we will supply benchmarking data of focused plenoptic cameras by synthetic ground truth datasets and high-quality real images captured under controlled conditions by Raytrix cameras.

Index Terms— Plenoptic, Lightfield, Multi-focus, Toolbox, Dataset, Micro-Lens Array, Plenoptic 2.0, Raytrix

1. INTRODUCTION

The recent growth in lightfield technologies combined with the latest research results has highlighted some major challenges within the lightfield community. From the definition itself, what is called lightfield and how different subtypes can be distinguished, to the evaluation of different approaches, where a lack of adequate material is shown.

While for binocular stereo images several different datasets covering a wide range of setup and applications have been proposed, this is not the case for lightfield data. A recent

*The work in this paper was funded from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

[†]The author performed the work while at Kiel University

analysis of stereo vision datasets in [1] takes 28 sets into consideration, the most famous being Middlebury [2] and KITTI [3], that alone reached more than 100 citation in the three main conferences (CVPR, ICCV, ECCV) in 2016 [1].

Lightfield datasets increased in the last year and different types of images are available, like synthetic images in [4], [5], real images taken with Lytro cameras [6], [7], [8] and images taken with a moving camera or an array of cameras in [9], but they show some important limitations. Firstly, most come without ground truth, as only [5] constitutes a real benchmark for numerical comparisons, and secondly they consist of the same type of images, not taking into account all possible lightfield imagery.

None or little effort has been focused on creating a dataset for plenoptic 2.0 images. Only one attempt was made in [10] to use OpenGL to emulate camera behaviour to obtain a plenoptic image of the Stanford Bunny, but no available dataset was produced. Introduced in [11], multi-focus plenoptic cameras deliver another solution to the challenge of capturing lightfield in a single shot, and especially now after the discontinuation of Lytro cameras, they retain importance in the plenoptic field.

2. PRIOR WORK AND CONTRIBUTION

Many approaches have been proposed for lightfield and its several applications; it is therefore difficult and out of the scope of this work to review the whole literature. The focus will be on depth estimation, core of several applications, especially in the case of the plenoptic 2.0 cameras, because it needs geometry estimation of the scene in order to use the refocusing properties typical for lightfield acquired scenes.

Recently a combined effort of many scientists in the field was made to evaluate and analyze different depth estimation approaches using the same dataset [12]. The chosen dataset, created with a Blender plug-in to render lightfield scenes, was based on the images acquired with plenoptic 1.0 cameras and consisted of 9x9 views. Other types of lightfields were not taken into consideration.

The works focusing on the plenoptic 2.0 cameras ranges

6. Publications

from calibration, [13], [14], [15], to spatial resolution in [16] to depth estimation. Dense depth maps were achieved through stereo matching in [17], while in [18] feature matching is used to obtain sparse depth maps that are successively filled. Recent works focus on optimization or mixing of lens pattern selections and the creation of synthetic images, either with Blender or OpenGL, as in [19] and [10].

Therefore our work addresses the challenge and provides a tool for multi-focus plenoptic cameras to contribute to the research of different lightfield acquisition methodologies.

The contribution of this work is twofold:

1. It provides a toolbox for plenoptic 2.0 images, that contains the first available dataset of such images, consisting of both real images taken with Raytrix cameras (where raw and processed images are available along with a configuration file) and synthetic images with relative ground truth and a completely open-source repository with the code used to work on these images;
2. It evaluates different methods to estimate depth from these kind of images using both synthetic numerical evaluation purposes and real images.

3. THE PLENOPTIC TOOLBOX

The Plenoptic Toolbox consists of open-source code that allows the development of several applications using plenoptic 2.0 images. All the source code is available online at the GitHub page [20], in the *python* language, for research purposes.

The provided code uses a dictionary to load and store the micro-images with their parameters. This allows for an efficient utilization and offers the flexibility to develop new methods and applications without the need of a new implementations.

3.1. The Dataset

Given the challenges related to the acquisition and the usage of multi-focus plenoptic cameras, few reliable images are available online: our work provides the first online database of plenoptic 2.0 images, taken with different cameras to guarantee a homogeneous distribution.

At the time of this publication, Raytrix R29 and R42 cameras were chosen because of their higher quality of the pictures. All pictures were taken under controlled conditions.

The creation of the dataset followed the plan to present different challenges in the depth estimation: as it's possible to see in the supplementary material, the acquired images present white background and thus textureless regions as in *Cards* or *Cars*, as well as textured background in *University* or *Dragon*, different types of specularities in *Specular*, fine and detailed structures in *Hawaii*, and slopes with texture for an easier matching in *Dixit*.

Along with the real images, a set of synthetic images is provided: they were created using Blender to emulate the micro-lens array grid, therefore in an ideal condition. No distortion have been applied, and they all have the corresponding ground truth.

The multi-focus properties of the cameras are also taken into account in the generation of the synthetic images, delivering micro-images with different amount of blur according to their focal lengths, to be as close as possible to the real scenes.

4. BENCHMARKING OF DEPTH ESTIMATION

Due to their recent introduction, there are not many technique forming the state-of-the-art for depth estimation on plenoptic 2.0 images. As a start, work from [19] and [17] is used, where depth is computed by stereo matching the micro-images using the well known semi-global matching method.

Five different similarity measures to compute the cost volume are analyzed: Absolute Difference (AD), Squared Difference (SD), Census, Gradient with AD and Normalized Cross Correlation. These measures are widely used and constitutes the basis for state-of-the-art methods, as in [22] where a learning approach is used to choose the best matching costs.

The existing overviews are mainly referred to binocular stereo [23] and does not take into account neither the multiview case or the nature of small micro-images, thus our evaluation extends these works for plenoptic images.

4.1. Evaluation

The evaluation procedure is divided into two parts, because of the nature of the different images. For the real images, only a visual subjective comparison can be made, as already highlighted in [18], due to the lack of ground truth. We provide some visual comparison between the different methods (more in the supplementary material) and a general overview.

In the case of synthetic images it's possible to analyze more in detail the results. Following the Middlebury stereo vision benchmark [2] and the latest works in this field as [5], [12] and [24], the most representative criteria for depth estimation classification has been reproduced.

The criteria to be used for benchmarking were chosen based on their significance. For example, with respect to binocular stereo vision, plenoptic 2.0 images present special properties: such images do not suffer from occlusion problem, so this criterium, of large importance in the stereo case, was discarded for our benchmark.

The errors have been calculated under the form of:

- Number of pixels that show a disparity error larger than ϵ , with $\epsilon \in [0, 2]$ are estimated. As a reference the two values for $\epsilon = 1, 2$ are chosen, like the so-called *bad 1.0* and *bad 2.0* of Middlebury).

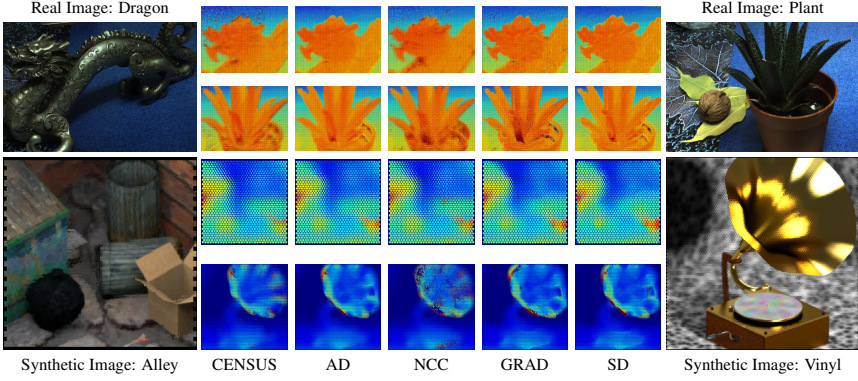


Fig. 1. Samples of images from the dataset with their respective estimated depth maps. *First row:* real images taken with R29 Raytrix camera, along with an excerpt of their depth map estimated using the similarity measures above described. *Second row:* synthetic images generated with Blender, along with their depth maps. The ground truth is not shown here. *AD* = absolute difference. *NCC* = normalized cross correlation. *GRAD* = gradient. *SD* = squared difference. Please refer to the colored version for a better visualization. Images are visible in the supplementary material and available at [21].

- Average Error (AE) and Mean Squared Error (MSE)

$$AE = \sum_{i,j \in I} \frac{|d_{i,j} - g_{i,j}|}{|I|} \quad MSE = \sum_{i,j \in I} \frac{(d_{i,j} - g_{i,j})^2}{|I|} \quad (1)$$

- Bumpiness measurement, that accounts for smoothness of estimation, using the formula from [5], changing the clamping threshold (B_{thresh}) according to our depth range.

$$B = \sum_{i,j \in I} \frac{\min(B_{thresh}, |d_{i,j} - g_{i,j}|)}{|I|} \quad (2)$$

where in the presented results $B_{thresh} = 0.25$, $d_{i,j}$ and $g_{i,j}$ represent respectively disparity estimated and ground truth at the i and j pixel, and I indicates the circle-shaped micro-image that is used for the calculations.

- Errors around depth discontinuities, similar to the criterion used in [24]. The image is divided into two parts, where one contains the pixels around edges in the disparity maps and the other one the rest. The edges were obtained through the OpenCV implementation of the Canny algorithm followed by a dilation operation to obtain the area around it (one pixel per side).

4.2. Results on Real Images

In this section some of the results are shown to back up the general considerations. Because of the limitations of synthetic images, real images still provide for a more challenging task, having to deal with physical lenses and sensors. Moreover, the variety of the scenes that can be captured allow the testing of the algorithms for different purposes, targeting specific issues.

The results show that the current algorithm, independently from the similarity measure chosen, fails in reconstructing large textureless surfaces. For this reason, two different sets of images were acquired: one with white textureless background, where the algorithm shows low quality results, and one with textured background, where it delivers robust estimation. This happens because of the local approach used and can be improved by choosing a global solution or by applying a post processing refinement step, for example a filling algorithm. This will be addressed in future research.

The similarity measures analyzed show quite some difference in the analyzed images. The AD and CENSUS obtain satisfactory results and high quality depths in textured scenes, while NCC shows alternate performances depending on the scene and large error that affect the whole image, so that its usage can be considered mainly in combination with other similarity measure. The depth maps obtained with SD and Gradient, lastly, show lower quality and larger areas with a wrong estimation.

6. Publications

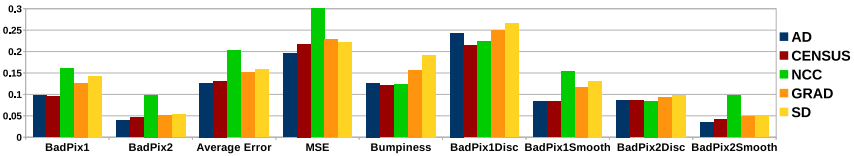


Fig. 2. The different criteria used for evaluation. Some of the values have been scaled for visualization purposes. *BadPix1,2* = Percentage of pixels which error exceeds 1,2 pixels. *Average Error* = Average of the absolute error in pixel. *MSE* = Mean squared Error. *Bumpiness* = Bumpiness measure taken from [5]. *..Disc* = .. Around depth discontinuities. *..Smooth* = .. Around smooth areas (not considered depth discontinuities.) Please refer to the colored version for a better visualization.

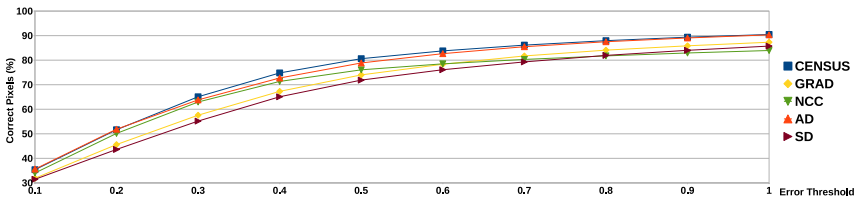


Fig. 3. For each similarity measure, the percentage of correctly estimated pixel on the synthetic scenes is plotted for the increasing error's thresholds on the x-axis. Please refer to the colored version for a better visualization.

4.3. Results on Synthetic Images

The synthetic images were analyzed on the above defined criteria: AD and CENSUS are the methods who achieve an overall higher quality, with NCC that shows some interesting characteristics and Gradient and SD that fail larger areas. On the average error measure and the Bad Pixel 1.0 / 2.0, the AD and CENSUS outperforms the other methods and achieve comparable results. Almost same case in the Mean Squared Error apart from a higher value for NCC. NCC has lower performances in the Bad Pixel 1.0 and 2.0, showing the highest number of errors with large value. In the Bumpiness criterium AD, CENSUS and NCC reach the same level and SD and GRAD have weaker performances.

The last criteria, indicating the number of erroneous pixels around depth discontinuities and in smoother areas, give a reference about the robustness of the estimation: NCC, for example, obtains a good score in the discontinuities areas, while performing poorly on smooth surfaces. Apart from the NCC case, the errors confirm that depth discontinuities are still the most challenging parts. This might be emphasized by the approach that does not include a refinement step for accurate reconstruction of fine structures.

Fig. 3 shows the behaviour of the number of correct pixels varying the error threshold. As expected, AD and CENSUS reach higher levels. The estimation using NCC has an unex-

pected curve: for low thresholds it maintains same level of AD and CENSUS, but makes higher amount of larger errors, resulting to lower performances in most of the measurements. This suggests that a combination of NCC with other measurements based on its confidence could lead to improvements.

5. CONCLUSION

The presented work extends the benchmarking for stereo vision to the plenoptic 2.0 images. It contributes to the development and spreading of such technologies, providing an important tool for future research; moreover, it makes available a dataset of images that is, up to our knowledge, the first of its kind. The dataset is meant to be continuously updated with new images for specific purposes and increasing difficulties.

This will not only allow an easier comparison of different methodologies of disparity estimation techniques, but also push other possible applications: in this direction, the next step will be to develop novel approaches for compression of such images.

Lastly, we provide a first version of plenoptic 2.0 benchmark, where different similarity measures are compared. Even though at the current state-of-the-art the changes are quite simple, it is to be seen as standard for an easy comparison for future development, a missing tool in this field.

6. REFERENCES

- [1] O. Zendel, K. Honauer, M. Murschitz, M. Humenberger, and G. F. Dom nguez, "Analyzing computer vision data - the good, the bad and the ugly," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [2] D. Scharstein, R. Szeliski, and H. Hirschmuller, "Middlebury stereo vision benchmark," <http://vision.middlebury.edu/stereo/>, [Online, 2018].
- [3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [4] G. Wetzstein, "Synthetic light field archive," <http://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php>, [Online, 2018].
- [5] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision (ACCV)*, 2016.
- [6] A. Mousnier, E. Vural, and C. Guillemot, "Lytro first generation dataset," <https://www.irisa.fr/temics/demos/lightField/index.html>, [Online, 2018].
- [7] A. Ghasemi, N. Afonso, and M. Vetterli, "Lcav-31: A dataset for light field object recognition," in *Proceedings of the SPIE, vol. 9020, International Society for Optics and Photonics*, 2014.
- [8] Caner Hazirbas, "4.5d lightfield-depth benchmark," <http://hazirbas.com/datasets/ddff12scene/>, [Online, 2018].
- [9] A. S. Raj, M. Lowney, and R. Shah, "Light-field database creation and depth estimation," <http://lightfield.stanford.edu/>, 2016.
- [10] R. Ferreira, J. Cunha, and N. Goncalves, "Multi-focus plenoptic simulator and lens pattern mixing for dense depth map estimation," in *EUROGRAPHICS*, 2016.
- [11] C. Perwaß and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Proceedings of SPIE - The International Society for Optical Engineering 8291-4*, 2012.
- [12] K. Honauer and O. Johannsen et al., "A taxonomy and evaluation of dense light field depth estimation algorithms," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [13] S. Nousias, F. Chadebecq, J. Pichat, P. Keane, S. Ourselin, and C. Bergeles, "Corner-based geometric calibration of multi-focus plenoptic cameras," in *International Conference on Computer Vision (ICCV)*, 2017.
- [14] Y. Bok, H.G. Jeon, and I.S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, February 2017.
- [15] C. Heinze, S. Spyropoulos, S. Hussmann, and C. Perwaß, "Automated robust metric calibration algorithm for multifocus plenoptic cameras," *IEEE Transactions on Instrumentation and Measurements*, May 2016.
- [16] M. Damghanian, R. Olsson, M. Sjöström, A. Erdmann, and C. Perwaß, "Spatial resolution in a multi-focus plenoptic camera," in *International Conference on Image Processing (ICIP)*, 2014.
- [17] O. Fleischmann and R. Koch, *Lens-Based Depth Estimation for Multi-focus Plenoptic Cameras*, pp. 410–420, Springer International Publishing, 2014.
- [18] R. Ferreira and N. Goncalves, "Fast and accurate micro lenses depth maps for multi-focus light field cameras," in *German Conference on Pattern Recognition (GCPR)*, 2016.
- [19] L. Palmieri and R. Koch, "Optimizing the lens selection process for multi-focus plenoptic cameras and numerical evaluation," in *2nd LF4CV Workshop. CVPR*, 2017.
- [20] L. Palmieri, "The plenoptic toolbox 2.0," <https://github.com/PlenopticToolbox/PlenopticToolbox2.0>, [Online, 2018].
- [21] L. Palmieri, "Multi-focus plenoptic images dataset," <https://drive.google.com/drive/folders/17I6nTf4GLYi09fdWITEy155F-OaonaeQ>, [Online, 2018].
- [22] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon, "Depth from a light field image with learning-based matching costs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.
- [23] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 9, pp. 1582–1599, 2009.
- [24] K. Honauer, L. Maier-Hein, and D. Kondermann, "The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms," in *International Conference on Computer Vision (ICCV)*, 2015.

6. Publications

6.3 Publication 3

Matching Light Field Datasets From Plenoptic Cameras 1.0 and 2.0

Waqas Ahmad, Luca Palmieri, Reinhard Koch and Mårten Sjöström

Published in

2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON) [APK+18]

© 2018 IEEE. Reprinted, with permission, from Waqas Ahmad

DOI: 10.1109/3DTV.2018.8478611

MATCHING LIGHT FIELD DATASETS FROM PLENOPTIC CAMERAS 1.0 AND 2.0

Waqas Ahmad¹, Luca Palmieri², Reinhard Koch², Mårten Sjöström¹

¹Department of Information Systems and Technology, Mid Sweden University, Sundsvall, Sweden

²Department of Computer Science, Christian-Albrechts-Universität, Kiel, Germany

ABSTRACT

The capturing of angular and spatial information of the scene using single camera is made possible by new emerging technology referred to as plenoptic camera. Both angular and spatial information, enable various post-processing applications, e.g. refocusing, synthetic aperture, super-resolution, and 3D scene reconstruction. In the past, multiple traditional cameras were used to capture the angular and spatial information of the scene. However, recently with the advancement in optical technology, plenoptic cameras have been introduced to capture the scene information. In a plenoptic camera, a lenslet array is placed between the main lens and the image sensor that allows multiplexing of the spatial and angular information onto a single image, also referred to as plenoptic image. The placement of the lenslet array relative to the main lens and the image sensor, results in two different optical designs of a plenoptic camera, also referred to as plenoptic 1.0 and plenoptic 2.0. In this work, we present a novel dataset captured with plenoptic 1.0 (Lytro Illum) and plenoptic 2.0 (Raytrix R29) cameras for the same scenes under the same conditions. The dataset provides the benchmark contents for various research and development activities for plenoptic images.

Index Terms— Plenoptic, Light-field, Dataset

1. INTRODUCTION

The seven-dimensional (7D) plenoptic function completely represents the light information within an observable space [1]. In the observable space, each light ray has a spatial position (3D), a direction (2D), a time instant (1D), and a wavelength that reflects the color (1D) information. However, the present technology has physical limitation to capture a light field using the 7D plenoptic function. In order to reduce the dimensions of the plenoptic function, a set of constraints are used for the scene. The time information is not required when the scene is assumed static. The wavelength is sampled using RGB channels, the wavelength parameter is constant for each color channel and can be omitted in the representation. Finally, an occluder free scene makes it possible to capture the incoming light rays onto a 2D plane instead of capturing light ray at each point of the space. Hence, the 7D plenop-

tic function is reduced to 4D when fixing the other parameters [2] and the captured spatial and angular information of the scene is referred to as light field (LF). The additional angular information has high significance since it enables various post-processing application, e.g. 3D scene reconstruction, refocusing at different depth planes, synthetic aperture, and digital zoom.

In order to record spatial and angular information, multiple traditional cameras have been mounted on a camera rig and scene is captured at a single time instant, e.g. in [2]. The light field captured with a multi-camera system is referred to as sparsely sampled light field, and the pursuit of having densely sampled light field introduces a new camera technology. The spatial and angular information of the scene is captured in the latter case using a single camera, also referred to as plenoptic camera. The idea of plenoptic capture was first introduced by Gabriel Lippmann in 1908 [3]. However, in 2006 the first commercial model was introduced by Ren Ng at Lytro [4]. In plenoptic camera, a lenslet array is introduced between main lens and image sensor that multiplex angular and spatial information onto a single image. Each microlens captures one position and multiple angular information and the model is referred to as plenoptic 1.0. In 2009, another version of the plenoptic camera also referred to as plenoptic 2.0 was proposed [5] with a slight change in optical design. Each micro-lens for such a plenoptic 2.0 camera captures a mixture of spatial and angular information of the scene. In 2012, based on plenoptic 2.0 camera model, Raytrix has introduced multi-focus plenoptic camera [6]. Micro-lenses with three different focal lengths were used to increase the depth of field of the captured LF.

Light field datasets are available online with different characteristics. Early contributions used multiple cameras as in [7] and [8], while more recent datasets contain LF images captured using the Lytro camera, as in [9], [10] and [11] for general light field processing applications. A dataset consists of synthetic contents [12] was also used as a benchmark for depth estimation schemes.

In recent past, plenoptic image processing has gained significant attention from the research community. Various competitions for plenoptic image compression were organized [13, 14] and also novel methods related to 3D scene reconstruction and depth estimation [15, 16] were proposed.

6. Publications

However, in most of the experiments plenoptic images captured with the Lytro camera were used due to the availability of Lytro datasets [9, 10]. In this paper, we present a novel dataset where the same scene under the same conditions was captured using Plenoptic 1.0 and Plenoptic 2.0, providing benchmark contents for LF applications and algorithms. The rest of the paper is organized as follows. The basic functionality of a plenoptic camera is explained in section 2. In section 3, the experimental setup is presented, and section 4 reports the contents of the dataset. The presented work is concluded in section 5.

2. THE PLENOPTIC CAMERA

2.1. Plenoptic camera 1.0

The plenoptic camera 1.0 has a micro-lens array (MLA) at the focal plane of the main lens as shown in Fig.1. The image behind each micro-lens contains the information about only one spatial point. The pixels in such a micro-lens image contain the angular information for the light passing this spatial position. The number of pixels in the micro-lens image defines the angular resolution, i.e. the number of different view points. The spatial resolution of the captured LF is determined by the number of micro-lenses.

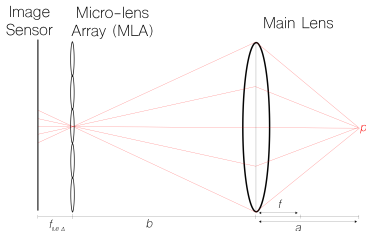


Fig. 1: A schematic configuration of the Plenoptic 1.0 camera: the MLA is placed at the focal plane of the main lens, and image sensor is placed at distance f_{MLA} (focal length of micro-lens). A point p is seen only from one micro-lens. The parameters correspond to the thin lens equation $\frac{1}{f} = \frac{1}{a} + \frac{1}{b}$.

2.2. Plenoptic camera 2.0

In plenoptic camera 2.0, the micro-lens array is focused onto the image plane of the main lens as shown in Fig.2. Each micro-lens records a part of the scene so that a point is visible across multiple micro-lenses from slightly different perspectives, generating micro-images with overlapping areas. The trade-off between the spatial and the angular resolution depends on the overlap between micro-images: lower overlap results in larger spatial resolution and vice versa.

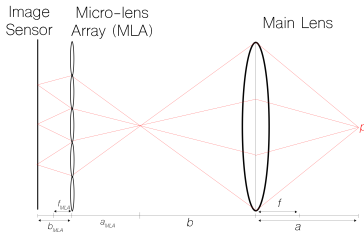


Fig. 2: A schematic configuration of the Plenoptic 2.0 camera: the MLA is focused on the focal plane of the main camera lens. A point p is seen by many micro-lenses. The parameters correspond to the thin lens equation $\frac{1}{f} = \frac{1}{a} + \frac{1}{b}$.

The multi-focus plenoptic camera produced by the Raytrix has an additional characteristic: the MLA contains three different lens types with different focal length. This extends the depth-of-field of the captured LF, but introduces other challenges in the manipulation of such images: each micro-image shows a different amount of defocus blur based on the depth of the scene.

3. EXPERIMENTAL SETUP

The dataset is captured using two different plenoptic cameras: Illum from Lytro and R29 from Raytrix. The former follows the plenoptic 1.0 model and the latter follows the plenoptic 2.0 model. Fig.3 shows a zoomed area of the scene captured with Illum and R29 cameras. The scenes selected for the dataset were captured under controlled conditions. Instead of using the natural light source, the scenes were captured in a closed room with two movable light sources. The cameras were mounted onto a multi-camera rig that was mechanically controlled to move the cameras with millimeter precision as shown in Fig.4. In this way, both cameras captured the scene from the same view point.

The R29 camera was selected as a reference and parameters of the Illum camera were adjusted accordingly. The focal distance of 0.8 meters was used to capture the dataset. Later on, zoom of the Illum camera was adjusted to match the field of view of the R29 camera. However, the Illum camera has a slightly higher vertical field of view compared to the R29 camera. In the R29 camera, the aperture size was adjusted according to the size of micro-lens aperture ($\frac{f}{8}$ in our case). The ISO parameter is also fixed in the R29 camera and only the shutter speed was adjusted to account for the exposure. The Illum camera has fixed aperture size ($\frac{f}{2}$) and the exposure was adjusted using ISO and shutter speed. To achieve the best quality in the captured images, the minimum ISO value (80) was used for the Illum camera. The shutter speed

6.3. Publication 3

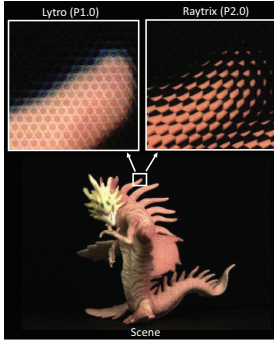


Fig. 3: A close up of the same scene captured by Lytro Illum and Raytrix R29 cameras, respectively top left and top right.



Fig. 4: The experimental setup used for capturing the proposed dataset. The Lytro Illum and Raytrix R29 plenoptic cameras were used to capture the same scene from same view point.

was manually adjusted with respect to each scene.

4. THE PLENOPTIC DATASET

The captured dataset is made publically available [17] for research community to design, test and benchmark various LF image processing algorithms. The dataset contains 31 LF images, captured with two different plenoptic cameras. A subset of LF images captured with Lytro Illum camera are shown in Fig.5 and their corresponding images captured with Raytrix R29 camera are shown in Fig.6. Keeping in view the content requirements for various applications, the dataset is captured in such a way that LF images inherit specific properties, e.g. different colors, objects at different depths, texture, shapes, and occlusion.

4.1. Lytro

The Lytro camera provides a calibration data file that contains the white image database along with camera specific

information. The white image database is pre-computed by the manufacturer for each camera. Each white image corresponds to single zoom step and focus step setting. The white image is used in pre-processing stage, e.g. deconvolution. The Lytro camera stores plenoptic image in Light field Raw (LFR) format. The captured LFR images can be processed using Matlab Lytro toolbox [18] to perform demosaicing and deconvolution. Moreover, the processed plenoptic image can be converted into sub-aperture representation.

4.2. Raytrix

The Raytrix application RxLive 4.1 [19] is used to capture the LF images. For each LF image the Raytrix dataset contains a calibration file (provide information about Raytrix camera parameters), a raw LF image (without debayering and demosaicing), a processed LF image (after debayering and demosaicing) and a total focus image.

5. CONCLUSION

The paper presents a novel and publicly available plenoptic image dataset. Each scene was captured by using two different plenoptic cameras, namely Illum from Lytro (based on plenoptic 1.0 model) and R29 from Raytrix (based on plenoptic 2.0 model). Both cameras were mounted onto a mechanically driven rig with millimeter precision and scenes were captured from a single view point. The captured LF images inherit various properties, e.g. objects at different depths, different colors, texture, shapes, and occlusions. The presented dataset provides benchmark contents for LF image processing algorithms, e.g. disparity estimation, compression, water marking, segmentation and etc. The detailed information about the presented dataset is available at [17].

6. ACKNOWLEDGEMENT

The work in this paper was funded from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

7. REFERENCES

- [1] E. H. Adelson and J. R. Bergen, “The plenoptic function and the elements of early vision,” in *In Computation Models of Visual Processing*, M. Landy and J.A. Movshon (Eds.). MIT Press, Cambridge, 1991, pp. 3–20.
- [2] M. Levoy and P. Hanrahan, “Light field rendering,” in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.

6. Publications

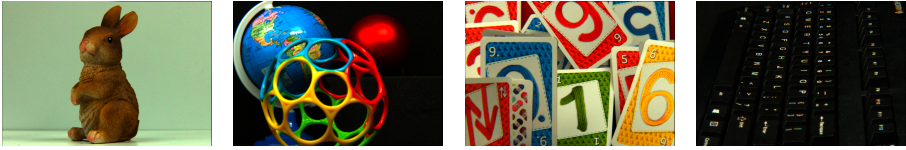


Fig. 5: A subset of LF images captured with Lytro Illum camera. The figure shows the central sub-aperture views.

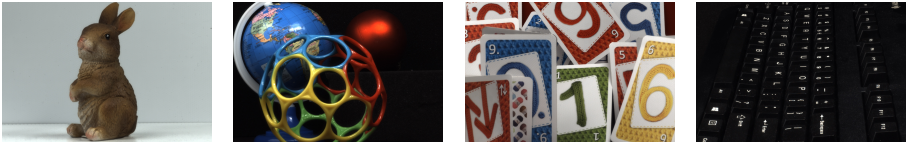


Fig. 6: The corresponding LF images captured with Raytrix R29 camera. The figure shows the total focus images.

- [3] G. Lippmann, “Epreuves reversibles donnant la sensation du relief,” *J. Phys. Theor. Appl.*, vol. 7, no. 1, pp. 821–825, 1908.
- [4] R. Ng, M. Levoy, B. Mathieu, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a handheld plenoptic camera,” *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.
- [5] A. Lumsdaine and T. Georgiev, “The focused plenoptic camera,” in *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009, pp. 1–8.
- [6] C. Perwass and L. Wietzke, “Single lens 3d-camera with extended depth-of-field,” in *Human Vision and Electronic Imaging XVII*. International Society for Optics and Photonics, 2012, vol. 8291, p. 829108.
- [7] The (New) Stanford Light Field Archive, “<http://lightfield.stanford.edu/lfs.html>,” .
- [8] Synthetic Light Field Archive (MIT), “<http://web.media.mit.edu/>”
- [9] M. Rerabekand T. Ebrahimi, “New light field image dataset,” in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, number EPFL-CONF-218363.
- [10] P. Paudyal, R. Olsson, M. Sjöström, F. Battisti, and M. Carli, “Smart: A light field image quality dataset,” in *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 2016, p. 49.
- [11] A. S. Raj, M. Lowney, and R. Shah, “Light-field database creation and depth estimation,” 2016.
- [12] K. Honauer, O. Johannsen, and B. Goldluecke D. Kondermann, “A dataset and evaluation methodology for depth estimation on 4d light fields,” in *Asian Conference on Computer Vision*. Springer, 2016, pp. 19–34.
- [13] Call for Proposals on Light Field Coding, “Jpeg pleno,” *ISO/IEC JTC 1/SC29/WG1N74014, 74th Meeting, Geneva, Switzerland*, January 15-20, 2017.
- [14] M. Rerabek, T. Bruylants, T. Ebrahimi, F. Pereira, and P. Schelkens, “Icme 2016 grand challenge: Light-field image compression,” *Call for proposals and evaluation procedure*, 2016.
- [15] M. Kim, T. Oh, and I. S. Kweon, “Cost-aware depth map estimation for lytro camera,” in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 36–40.
- [16] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, “Depth from combining defocus and correspondence using light-field cameras,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 673–680.
- [17] Dataset, “<https://doi.org/10.6084/m9.figshare.6115487>,” .
- [18] D. G. Dansereau, O. Pizarro, and S. B. Williams, “Decoding, calibration and rectification for lenselet-based plenoptic cameras,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1027–1034.
- [19] Raytrix RxLive 4.1, “<https://raytrix.de/downloads/>, (accessed: 2018-03-02),” .

6.4 Publication 4

Simulation of Plenoptic Cameras

Tim Michels, Arne Petersen, Luca Palmieri and Reinhard Koch

Published in

2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON) [MPP+18]

© 2018 IEEE. Reprinted, with permission, from Tim Michels

DOI: 10.1109/3DTV.2018.8478432

6. Publications

SIMULATION OF PLENOPTIC CAMERAS

Tim Michels, Arne Petersen, Luca Palmieri, Reinhard Koch

CAU Kiel, Department of Computer Science

ABSTRACT

Plenoptic cameras enable the capturing of spatial as well as angular color information which can be used for various applications among which are image refocusing and depth calculations. However, these cameras are expensive and research in this area currently lacks data for ground truth comparisons. In this work we describe a flexible, easy-to-use Blender model for the different plenoptic camera types which is on the one hand able to provide the ground truth data for research and on the other hand allows an inexpensive assessment of the cameras usefulness for the desired applications. Furthermore we show that the rendering results exhibit the same image degradation effects as real cameras and make our simulation publicly available.

Index Terms — Light Field, Plenoptic Camera, Simulation, Blender

1. INTRODUCTION

Tracing back to the ideas of Lippmann [1] and Ives [2] the concept of capturing light fields with a single camera has regained interest during the past decade due to its commercially available realizations in the form of plenoptic cameras by Lytro [3] and Raytrix [4]. Depending on the model, these cameras can be rather expensive, thus an accurate simulation of plenoptic cameras could enable a cheap and uncomplicated assessment of the usefulness for different applications. Furthermore, major parts of the research related to plenoptic cameras are focused on their calibration and the reconstruction of depth images as well as color images with a modified depth of field (DoF) or altered viewpoint. These would greatly benefit from realistic, simulated ground truth data. However, the simulation of realistic plenoptic camera data is non-trivial due to the setup of these cameras. The basic concept of the two types of plenoptic cameras shown in Fig. 1 and Fig. 2 is the use of a microlens array (MLA) between a conventional camera's main lens and its image sensor in order to capture not only position dependent but also view angle dependent information. Accordingly every light ray that reaches the sensor of a plenoptic camera has passed through the main lens system and one microlens and thus is affected by the properties of both.

While several related works on light field imaging focus on camera array data, the works that actually include or explicitly describe the simulation of plenoptic cameras usually ignore some part of the multi-lens setup leading to unrealistically perfect data. Fleischmann et al. [5] synthesize plenoptic camera data without using a main lens, thus reducing the setup to a simple multi-camera array. Zhang et al. [6] and Liang et al. [7] use a simplified thin main lens, which does not lead to the degradation effects visible in real images, and Liu et al. [8] require the captured scene objects to be at an unrealistically large distance from the camera. In addition most of these previous works use forward ray tracing and restrict themselves to scene objects with simple geometries and Lambertian surfaces.

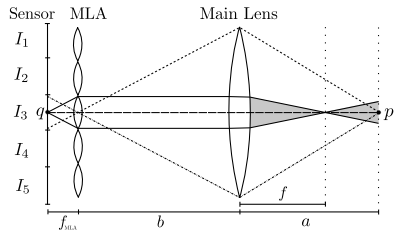


Figure 1: Plenoptic 1.0 camera as introduced by Adelson and Wang [10] and implemented by Ng [11]: The MLA is focused at infinity. Given $\frac{1}{f} = \frac{1}{a} + \frac{1}{b}$ the scene point p is seen by multiple pixels of the microlens image I_3 from slightly different angles. These pixels, however, also see a certain area around p as exemplarily shown for the pixel q , which results in a high angular but low spatial resolution [12].

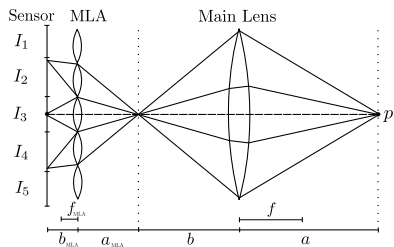


Figure 2: Plenoptic 2.0 camera as introduced by Lumsdaine and Georgiev [13]: The MLA is focused on the virtual image of the main lens. Given $\frac{1}{f} = \frac{1}{a} + \frac{1}{b}$ for the main lens as well as the microlenses, the scene point p is seen by multiple microlenses, but only one pixel per microlens image. This results in higher spatial but lower angular resolution compared to plenoptic 1.0 cameras [12].

Our contribution is a physically-based simulation of plenoptic 1.0 and 2.0 cameras in Blender [9] which includes a realistic model of the main lens as well as a configurable MLA and is made publicly available¹. Furthermore we analyze our synthesized images and show that these exhibit similar geometric and photometric degradation effects as images from Raytrix or Lytro cameras.

2. LENS EFFECTS IN PLENOPTIC CAMERAS

In this section we will give a description of the photometric and geometric degradation effects in plenoptic cameras. Due to the

¹<https://github.com/Arne-Petersen/Plenoptic-Simulation>

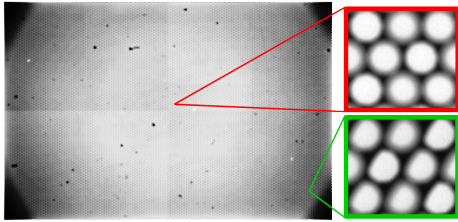


Figure 3: Vignetting effects: The left picture shows a heavily contrast enhanced image from a Raytrix R29 capturing an evenly illuminated white plane. Bright and dark spots are a result of microlens imperfections or dust particles on the MLA in combination with the contrast enhancement. The right pictures are unaltered sections of the original image showing the effect of the main lens aperture on the microlens images.

combination of the main and microlenses the effects that are observable in standard cameras are also present in plenoptic cameras, but have a different impact on the image quality. In the following we list major degradation effects and the problems arising in synthesizing them after rendering a defect-free multi-camera image without explicitly modeling the lenses of a plenoptic camera.

Radial distortion affects the main lens as well as the microlenses. However, the radial distortion of the microlenses is neither significant, because of the low resolution of the microlens images, nor efficient to handle due to the high number of microlenses. On the contrary, the radial distortion of the main lens plays a significant role in the plenoptic imaging process [14]. Synthesizing this effect via a simple inverse application of the radial distortion polynomial would result in a shift of the microlens images on the sensor leading to incorrect correspondences between a microlens area on the sensor and its image position.

Vignetting in the final image results from the addition of vignetting effects from the main lens and microlenses as shown in Fig. 3. The main lens vignetting, influenced by its aperture, does not only affect the amount of light reaching the microlenses, it also defines the shape of a microlens image. Furthermore, with increasing distance from the main lens optical axis, the microlens images are cut off to one side (compare Fig. 3) due to a limited exit pupil. These effects of the main lens vignetting are further amplified by the microlens aperture which causes additional vignetting. Despite the correction of this vignetting being simple via classical white image division [15], the complexity of the combined effect poses a problem for the feasibility of its synthesis. In order to generate the correct vignetting for a certain camera configuration a complex model is needed taking into account the main lens aperture configuration as well as the microlens position with respect to the main lens optical axis.

Depth distortion describes the influence of the microlens distance to the main lens optical axis on depth reconstruction algorithms. With increasing distance the depth error increases as a result of the so-called Petzval field curvature [14]. In order to synthesize this effect in a perfect image for a certain objective a model similar to the radial distortion has to be applied. This however requires precise knowledge on the Petzval field curvature of the given lenses.

Coma and *astigmatism* are further effects which, like the Petzval field curvature, can influence the focal properties of the lens and thereby the depth reconstruction. Using modern objectives their effects on the final image are usually negligible and accordingly

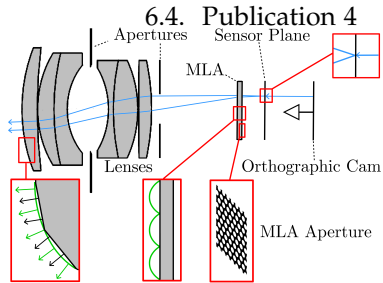


Figure 4: Overview of our plenoptic camera model in Blender including a comparison of real geometry (green) and the approximation in Blender (black). The blue lines show exemplary paths for rays casted from a camera pixel into the scene, i.e. the resulting color value of each ray is a sample value for the starting pixel.

it is only necessary to synthesize these effects if the main lens is known to exhibit a significant amount of these distortions. Then, however, the synthesis poses the same problem as the Petzval field curvature, namely the necessity of measuring these effects in order to build a precise distortion model.

Apart from these degradation effects the main lens also alters the scene geometry seen by the MLA. The virtual scene, given by projecting the real scene through the main lens, is a non-linearly scaled version of the real scene. While the geometry of the virtual scene could approximately be calculated by applying the thin or thick lens equation to the real scene, the textures and especially the lighting are hard to synthesize since straight rays of the real scene, as usually used in ray tracing applications, are projected to curves in the virtual image. Accordingly the use of a main lens model significantly reduces the complexity of the virtual scene formation.

Further types of degradation are given by the imperfections of real cameras such as lens material defects, dead pixels, inaccurately mounted objectives or simply dust particles on lenses or sensors. Depending on the extent of the defect or inaccuracy the effect ranges from the degradation of single microlenses (see Fig. 3) to complex *caustics* or a *tangential distortion* of the whole image due to a main lens tilt. Like most of the previously mentioned effects, these can also be synthesized with varying expenditure. Nevertheless, a combination of all or a subset of these effects would require several and in some combinations even more complex distortion models since some of the effects depend on each other. Therefore the easiest solution to overcome the necessity of unfeasible distortion models is the direct use of accurately modeled main and microlenses as described in the following section.

3. BLENDER MODELING

The use of Blender for modeling a plenoptic camera has several advantages. As a free and widely used software it enables every interested person to easily create realistic images by using the implemented Cycles render engine. This renderer is physically-based and supports different path tracing variants which allow object models with refractive materials to exhibit nearly the same effects as their real pendants. Therefore Blender is not only suitable to simulate multiple lenses but, in contrast to previous works' simulations, also allows the use of complex scenes including non-Lambertian and refractive materials as well as complex lighting

6. Publications

setups. However, Blender also has some limitations regarding the overall number of vertices and minimal correctly simulatable size of objects and distances between them, posing different challenges for the modeling of a plenoptic camera.

While the objective’s lenses can theoretically be modeled as sphere intersections, the limitation in the number of vertices leads to lens models with only approximately round surfaces. When such a model and its surface normals are used for the rendering, the final image shows triangle shaped artifacts resulting from the discretization of the surface and its normals. However, since the correct lens geometry can be seen as a combination of the approximating 3D model and additional thin lenses added to its surface (see Fig. 4), it suffices to only calculate the correct surface normals in order to simulate the expected lens behavior. A geometry correction is not necessary because thin lenses can be approximated by simple refracting planes. In other words, slightly incorrect surface normals distort the projection of rays through lenses significantly more than marginal surface displacements. Therefore we simply use the lens models material shader to calculate the correct surface normals and overcome this discretization issue.

In addition to the lenses and the objective’s aperture we add another aperture at the objective’s exit (see Fig. 4) to simulate the limited exit pupil as discussed in the context of vignetting.

Since the microlens effects are mostly negligible compared to the main lens effects it seems natural to render each microlens image separately by using the internal Blender camera and shifting it to the next microlens position afterwards. However, the possible Blender camera settings are bounded below in respect of focal distance, sensor size and DoF values. This leads to a restricted MLA configuration and e.g. prevents realistic results when the MLA is placed between the main lens and the virtual image which is a common setup for plenoptic 2.0 cameras. Therefore it is necessary to explicitly model the MLA as well as the image sensor. Because of the limitation regarding the total number of vertices it is impossible for the microlenses to be realistically modeled since MLAs usually consist of up to $2 \cdot 10^5$ microlenses each of which would need a smooth, round surface. Since we are mainly interested in creating microlens models with the correct focal length, we can use the lensmaker’s equation, which states the exact relation between focal length, radius of the front surface of the microlens and the IOR, in order to find a simpler MLA model. For a lens with IOR n , front surface curvature radius R_1 , a small thickness $d \approx 0$ and a flat back surface, i.e. $R_2 = \infty$, as shown in Fig. 4, we get

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n - 1)d}{nR_1R_2} \right) = \frac{n - 1}{R_1}.$$

Accordingly, the same focal length f can be achieved with every thin lens satisfying $R_1/(n - 1) = f$, especially lenses with nearly flat front surfaces and high IOR. As previously mentioned, the normals and index of refraction (IOR) are more important for the correct refraction of a ray than the exact surface geometry. Therefore a microlens with a nearly flat front surface can be approximated by a flat lens with recalculated normals. Hence we use a simple two plane model with high IOR for the MLA and calculate the correct normals for the nearly flat lenses in the MLA model’s material shader. Furthermore, we mask the back surface of the MLA to simulate the microlens apertures (compare Fig. 4). Finally the image sensor of the plenoptic camera is simulated as a combination of a simple plane equipped with a refractive shader and an orthographic Blender camera which is viewing the plane and is rendered via backwards path tracing as implemented in Cycles. While a real camera sensor pixel has a certain FOV and collects light from this range of directions, a perfectly refractive

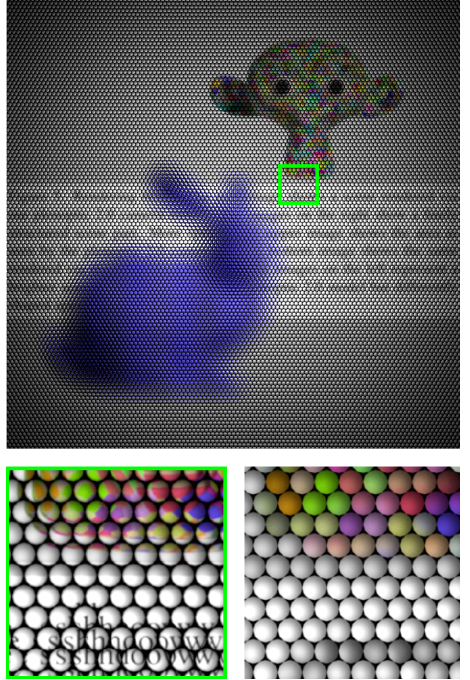


Figure 5: Top: A scene containing the Stanford bunny, a text plane and the Blender monkey rendered by our plenoptic 2.0 camera model. Bottom: Comparison of a finely structured scene part rendered via the plenoptic 2.0 (left) and 1.0 (right) model.

plane refracts a camera viewing ray into only one exact direction and therefore the corresponding orthographic camera pixel only sees a fraction of the light reaching the sensor plane. Thus, roughness is added to the refraction shader in order to allow the camera viewing rays for one pixel to be refracted randomly within a range of slightly different directions and therefore accumulating a realistic amount of light.

4. EVALUATION

For our tests we constructed a 100mm objective according to the double Gaussian lens model described in [16]. Furthermore we used microlenses with a focal length of 2mm and a diameter of 0.217mm. For the plenoptic 2.0 camera setup, the MLA distance to the objectives center was set to 123.3mm and the distance between MLA and sensor plane to 1.7mm. Consequently the microlens focal points are located slightly behind the sensor plane or, from a different point of view, the MLA is not focused at infinity but to a distance of 11.33mm according to the thin lens equation. For the plenoptic 1.0 setup the sensor plane is placed exactly 2mm behind the MLA, thus setting its focal distance to infinity.

Here we would like to remark, that our results slightly differ from the images captured with real Raytrix cameras as shown in Fig. 3 since we only use one microlens type and the modeled objective is

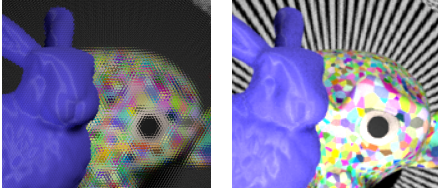


Figure 6: Cutout of a pleoptic 1.0 rendering and the corresponding section of one sub-aperture image.

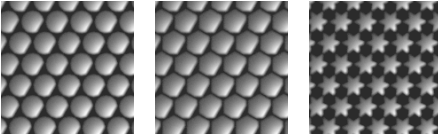


Figure 7: Vignetting effects in microlens images resulting from differently shaped (12-blade, 6-blade and star-shaped) main lens apertures and a limited exit pupil.

not equal to the real 100mm objective used for capturing that image. Moreover the extent to which the previously discussed effects are observable heavily depends on the objective. In our case the modeled objective exhibits radial as well as depth distortion to an extent that is only measurable but has nearly no visible effect on the renderings. These distortions, however, are inherent properties of lenses simulated via ray tracing thus there is no need for further validation regarding this aspect. Nevertheless, the rendering results show some other effects that are not only measurable but also clearly visible. As shown in Fig. 5, the renderings from the different pleoptic camera types exhibit the expected differences with respect to the trade-off between angular and spatial resolution for objects at a distance a that approximately satisfies the thin lens equation as shown in Fig. 1 and Fig. 2. While the pleoptic 2.0 images preserve fine details of the scene, the corresponding pleoptic 1.0 microlens images show a significant higher amount of blur. These correct imaging properties regarding the geometry can be further verified by the sub-aperture images (as exemplarily shown in Fig. 6) which exhibit the expected occlusion and reflection behavior. Here, due to the limited space, we refer the reader to our aforementioned repository, where additional renderings are available that show these effects.

Finally, the expected vignetting is also clearly observable. The rendered images show the main lens vignetting in regards of decreasing brightness towards the sensor edges (see Fig. 5) as well as regarding the shape and cut-off effect of the single microlens images (see Fig. 7).

5. CONCLUSION

While our model simulates pleoptic camera data quite realistically, there is still some room for improvement. Images of real (pleoptic) cameras exhibit chromatic aberrations and artifacts related to debayering as well as imperfections in the microlenses. Furthermore a variety of different main lens models could be implemented in order to simulate more complex and more recent objectives than the simple double Gaussian lens model we used. Nevertheless, with our simulation we present a useful basis for

6.4. Publication 4

future research that already covers the majority of the otherwise hard to synthesize effects.

Acknowledgment

This work was supported by the German Research Foundation, DFG, No. K02044/8-1 and the EU's Horizon 2020 program under the Marie Skłodowska-Curie grant agreement No 676401.

6. REFERENCES

- [1] G. Lippmann, "Epreuves reversibles, photographies integrales," *Academie des sciences*, 446451, 1908.
- [2] H. E. Ives, "A camera for making parallax panoramagrams," *JOSA*, vol. 17, no. 6, pp. 435–439, 1928.
- [3] Lytro inc. [Online]. Available: <https://www.lytro.com/>
- [4] Raytrix gmbh. [Online]. Available: <https://www.raytrix.de/>
- [5] O. Fleischmann and R. Koch, "Lens-based depth estimation for multi-focus pleoptic cameras," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 410–420.
- [6] R. Zhang, P. Liu, D. Liu, and G. Su, "Reconstruction of refocusing and all-in-focus images based on forward simulation model of pleoptic camera," *Optics Communications*, vol. 357, pp. 1–6, 2015.
- [7] C.-K. Liang and R. Ramamoorthi, "A light transport framework for lenslet light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, p. 16, 2015.
- [8] B. Liu, Y. Yuan, S. Li, Y. Shuai, and H.-P. Tan, "Simulation of light-field camera imaging based on ray splitting monte carlo method," *Optics communications*, vol. 355, pp. 15–26, 2015.
- [9] Blender. [Online]. Available: <https://www.blender.org/>
- [10] E. H. Adelson and J. Y. Wang, "Single lens stereo with a pleoptic camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 99–106, 1992.
- [11] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a handheld pleoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.
- [12] C. Perwass and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII*, vol. 8291. International Society for Optics and Photonics, 2012, p. 829108.
- [13] A. Lumsdaine and T. Georgiev, "The focused pleoptic camera," in *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009, pp. 1–8.
- [14] O. Johannsen, C. Heinze, B. Goldluecke, and C. Perwass, "On the calibration of focused pleoptic cameras," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013, pp. 302–317.
- [15] W. Yu, "Practical anti-vignetting methods for digital cameras," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 4, pp. 975–983, 2004.
- [16] C. Kolb, D. Mitchell, and P. Hanrahan, "A realistic camera model for computer graphics," in *Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '95. New York, NY, USA: ACM, 1995, pp. 317–324. [Online]. Available: <http://doi.acm.org/10.1145/218380.218463>

6. Publications

6.5 Publication 5

Robust Depth Estimation for Light Field Microscopy

Luca Palmieri, Gabriele Scrofani, Nicolò Incardona, Genaro Saavedra, Manuel Martínez-Corral and Reinhard Koch

Published in

Sensors 2019, 19(3), 500 in the Special Issue *Depth Sensors and 3D Vision*
[PSI+19]

DOI: 10.3390/s19030500



Article

Robust Depth Estimation for Light Field Microscopy

Luca Palmieri ^{1,*}, Gabriele Scrofanì ², Nicolò Incardona ², Genaro Saavedra ²,
Manuel Martínez-Corral ² and Reinhard Koch ¹

¹ Department of Computer Science, Christian-Albrecht-University, 24118 Kiel, Germany; rk@informatik.uni-kiel.de

² Department of Optics, University of Valencia, E-46100 Burjassot, Spain; gabriele.scrofanì@uv.es (G.S.); nicolo.incardona@uv.es (N.I.); genaro.saavedra@uv.es (G.Sa.); manuel.martinez@uv.es (M.M.-C.)

* Correspondence: lpa@informatik.uni-kiel.de; Tel.: +49 431 880-4843

Received: 18 December 2018; Accepted: 22 January 2019; Published: date

Abstract: Light field technologies have seen a rise in recent years and microscopy is a field where such technology has had a deep impact. The possibility to provide spatial and angular information at the same time and in a single shot brings several advantages and allows for new applications. A common goal in these applications is the calculation of a depth map to reconstruct the three-dimensional geometry of the scene. Many approaches are applicable, but most of them cannot achieve high accuracy because of the nature of such images: biological samples are usually poor in features and do not exhibit sharp colors like natural scene. Due to such conditions, standard approaches result in noisy depth maps. In this work, a robust approach is proposed where accurate depth maps can be produced exploiting the information recorded in the light field, in particular, images produced with Fourier integral Microscope. The proposed approach can be divided into three main parts. Initially, it creates two cost volumes using different focal cues, namely correspondences and defocus. Secondly, it applies filtering methods that exploit multi-scale and super-pixels cost aggregation to reduce noise and enhance the accuracy. Finally, it merges the two cost volumes and extracts a depth map through multi-label optimization.

Keywords: depth estimation; light field; microscope; stereo matching; defocus

1. Introduction

Light field microscopy was first introduced at Stanford in 2006 [1], and later improved in the same laboratory [2–4]. It consists of placing a microlens array (MLA) at the image plane of a conventional microscope, allowing for the capture of light field that records simultaneously both angular and spatial information of microscopic samples.

The main limitation of light field in microscopy is the spatial resolution [5]. To overcome this problem, a change of paradigm was necessary, so that the MLA is set, not at the image plane, but at the Fourier plane [6,7]. This realization of light field concept was named as Fourier integral Microscopy (FiMic). Light field microscopy has been used for several applications, such as brain imaging of neural activities in [8–10].

A common goal in microscopy is to estimate the three-dimensional structure of the observed sample. Industrial solutions reach a high accuracy of the reconstruction using different techniques as confocal microscopy [11], interferometry or variational focus [12], scanning electron microscopy (SEM) [13], optical profilometry [14,15], or stereo cameras [16].

However, these methods also present disadvantages in terms of real-time feasibility, sample preparation and costs. Light field concept provides a simpler and inexpensive approach that allows for many applications in real-time. On the other hand, because of the lack of texture and the presence

6. Publications

of repetitive patterns that characterize microscopic samples, the task of extracting a depth map from such light fields presents many challenges.

To address these challenges, different methods have been proposed, whose main limitation still is the final resolution: in [17] optical flow and triangular meshes are used, and in [18] a Lytro consumer camera is used to build a light field microscope and a variational multi-scale optical flow algorithm is used to estimate the depth. An interesting approach to estimate depth of thin structure has been proposed in [19], obtaining high quality results at the price of targeting only a small subset of biological images.

Depth estimation from light field images has already been largely studied and different approaches were proposed, using epipolar plane images [20,21], angular or spatial information [22,23], focal stack [24] and correspondences cues combined [25], explicitly modeled occlusion-aware approaches [26], robust pseudo random field [27] and learning based costs [28].

In previous works we extended similar applications to the case of focused plenoptic cameras (plenoptic 2.0) using MLA with a large number of microlenses, where the raw image contains micro-images capturing a portion of the scene. The main targets were the lens selection process for such MLA-based cameras [29] and evaluating different methods for depth estimation [30].

However, FiMic images constitute another special case. Because of MLA position and structure, the light field is sampled differently. It samples the light field as a conventional plenoptic camera (plenoptic 1.0), while at the same time the perspective views are arranged on a hexagonal grid and exhibit large disparities between them, thus an interpolation to transform them into a rectangular grid would produce strong artifacts.

The main contribution of our work is the creation of a framework where different methods are unified and adapted to obtain a more robust and accurate approach. It takes the above-mentioned conditions into account and provides a method for recovering accurate depth information using a sparser light field, i.e., lower number of views with higher disparity shift. The method can be applied not only to FiMic, but also to images acquired with conventional light field cameras, by choosing only a subset of the perspective views, or to sparser light fields.

The structure of the rest of the paper is as follows: in Section 2, the Fourier Integral Microscope will be described; in Section 3, the depth estimation workflow is explained; in Section 4, a Comparative Performance Analysis to prove the quality of the proposed method is performed; then in Section 5, a potential application is shown to enhance the importance of the contributions; and finally in Section 6, a brief summary of the proposed work is given.

2. Fourier Integral Microscope

The presented work takes as input light field microscopy images. The light field microscope used here is the FiMic described in [7]. In the FiMic design shown in Figure 1 the MLA is conjugated with the aperture stop (AS) of the microscope objective (MO). In this way, the sensor, that is placed at the focal plane of the MLA, captures the perspective views directly.

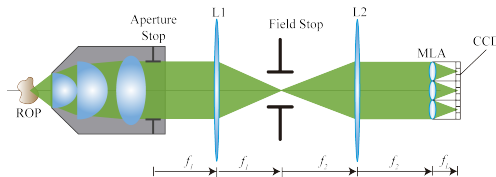


Figure 1. This is the schematic of the FiMic design. From left to right it is possible to distinguish the object, the microscope objective, two lenses (L1 and L2), one field stop, the MLA and the CCD sensor.

Changing the focal length of the two lenses L1 and L2, the designer can change the number of microlenses that fit in the diagonal of the AS. This affects the resolution limit (r) and the depth of field (DOF) as follows [7]:

$$r = M_n \frac{\lambda}{2NA} \quad (1)$$

$$DOF = \frac{5}{4} \frac{\lambda}{NA^2} M_n^2. \quad (2)$$

In these equations λ is the wavelength of the incoming light, NA is the numerical aperture of the MO and M_n is the number of microlenses that fit in the diameter of the AS. Note that usually the resolution capacity of an optical system is evaluated in terms of r^{-1} . It must be underlined that increasing M_n has two effects; a decrease of the resolution capacity and an increase of the DOF . An optimal setup for 3D microscopy aims to reach the highest resolution and the largest DOF , but as previously stated, these two factors are inversely affected by M_n . Therefore, depending on the sample to be reconstructed, M_n will be chosen according to the resolution and DOF required.

A typical example of an image acquired with this microscope can be seen in Figure 2, which includes seven elemental images, i.e., perspective views. These images can be seen as a particular case of a multi-view stereo imaging system, where different viewpoints are arranged on an hexagonal grid, having images aligned along at least three epipolar lines. This makes it suitable for correspondence matching.

Moreover, from such views a focal stack of refocused images can be extracted, by overlapping to the central one a shifted version of the elemental images, where the shift has to be oriented towards the center and will be the measure of the depth of the focal plane obtained. The depth of refocusing (z_R), with respect to the focal plane of the MO (considered as $z_R = 0\mu\text{m}$) can be calculated with the following formula [7]:

$$z_R = s \frac{f_{MO}^2}{f_L} \left(\frac{f_2}{f_1} \right)^2 \frac{\delta}{p}. \quad (3)$$

Here f_{MO} , f_L , f_2 and f_1 are respectively the focal lengths of the MO, MLA, L2 and L1. Besides, δ is the pixel's pitch, p is the MLA's pitch, and s is the integer number of shifted pixels applied to the focal stack. The use of defocus cue is then directly applicable to the generated focal stack. Due to these basic considerations, our approach aims to create a depth map by combining these two types of vision cues that are suitable for light field images.

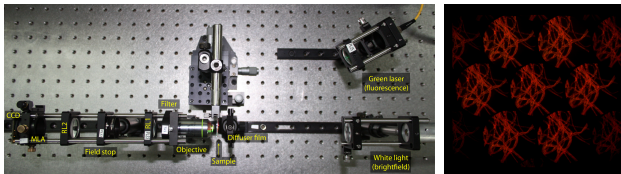


Figure 2. The setup used for the acquisition with the fluorescence laser used to illuminate the samples, and a sample output image acquired with such setup, from where the seven elemental images are visible.

3. Depth Map Calculation

The presented work builds on several successful ideas proposed for the depth estimation. The core consists in combining different visual cues, namely focus and correspondences, as in [25], to create a more versatile method, but it differs in the depth estimation.

The main contribution of this section is the combination of existing ideas and novel implementations. As it is possible to see in Figure 3, the depth estimation process can be divided into

6. Publications

intermediate steps that contributes to the outcome of the algorithm. To achieve high quality results without losing in robustness and flexibility, we designed the estimation process to allow full controls over the parameters and fine-tuning for each single step. The implementation is available at [25].

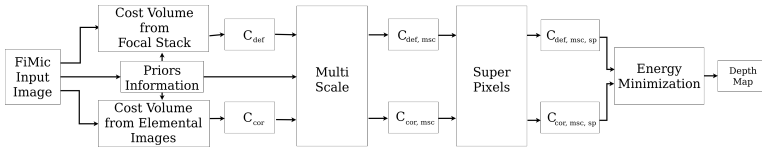


Figure 3. Pipeline of the depth estimation process. The name of cost volume (e.g., C_{def} , C_{cor}) are consistent with the ones used in the paper.

The pipeline consists in several steps: first, two cost volume cubes are calculated using respectively the elemental images, in Section 3.2, and the focal stack, in Section 3.3. These two cubes are then refined using a multi-scale approach similar to [31] in Section 3.4 and a contribution from superpixels inspired from [32] in Section 3.5. In Appendix A an overview of the parameters is given in Table A1.

Moreover, we introduce additional steps to increase the robustness of the algorithm and to adapt to a vast variety of input images. In Section 3.1 priors are incorporated in the fusion of the data under three different forms. A matting mask is applied to address the dark regions of the scenes, appearing mostly in transparent biological samples, where the empty areas do not capture light. Areas with high and low frequencies are weighted differently on high or low resolution depth estimation, and a failure prediction map is used to weight the contributions of the stereo matching along different epipolar lines.

The next step consists in fusing the two cost volumes and extracting a depth map. This is approached in Section 3.6 as an energy minimization problem, building an energy function that combines both cost volumes and finding the minimum using a multi-label optimization framework.

Finally, we used a post-processing filter to improve the estimation and obtain a smoother depth map. We applied a weighted median filter with integer weights followed by a guided filter, using in both cases the central image as a reference image.

3.1. Priors Information

Priors information have been incorporated into the depth estimation pipeline to deal with different kinds of input images, increasing the robustness and the reliability of the proposed approach.

3.1.1. Frequency Mapping

As explored in the literature in [33,34], the best improvements for the multi-scale approach are visible across areas with different frequencies: in high-frequencies areas high-resolution images can obtain the best results, while in low-textured and low-frequency areas a lower resolution leads towards more robust estimation.

Based on this consideration, we built a frequency map image using the difference of gaussians. The algorithm, calculating the difference in the amount of blur between two different blurred versions of the same image, is able to compute a frequency value for each pixel of the reference image.

$$F_{map} = I * (g(\sigma_1) - g(\sigma_2)) \quad (4)$$

where the reference image is denoted as I and $g(\sigma_1)$, $g(\sigma_2)$ are gaussian kernels of variance σ_1 and σ_2 . To obtain the desired result, we must ensure $\sigma_1 \neq \sigma_2$. The pixel values will then be used to weight the sum of the depth estimation at different scales, by quantizing the images into F_s levels, where each level corresponds to one scaled version.

3.1.2. Failure Prediction Map

In the stereo matching case, the correspondences search is conducted along the epipolar line: this inevitably leads to error in estimation of structures along that line, as pointed out in [35].

We built three different sobel-like kernel filters to detect edges at respectively 0, 60 and 120 degrees along the three epipolar lines of the hexagonal grid structures. By convolving the image with such a filter we obtain another weight map that favours the contribution of the estimation coming from the most appropriate direction. In Figure 4 it is possible to see the three kernels and an example of failure maps.

3.1.3. Matting

Some of the assumptions for natural images are not valid when dealing with microscope images. Because of the nature of the object and fluorescence illumination, some scenes exhibit a uniformly dark colored background.

In such dark and uniformly colored areas, almost every possible approach for depth estimation is doomed to fail. Based on these considerations, we adapt a matting approach to mask the areas we do not want to analyze. The proposed method has two steps: the first is creating a so-called trimap, that consists in heavily quantizing the image to three levels representing respectively foreground, unknown areas and background. This can be done by applying a multi-level thresholding on the focused image.

After this trimap is built, the next step consists in determining if the unknown areas belong to the foreground or background. We do this by applying one of the top performer methods in this areas, the three-layer graph approach in [36]. It introduces a new measure for discriminating pixels by using non-local constraint in the form of a three layer graph model to supplement local constraint, consisting of color line model, forming a quadratic programming that can be solved to obtain the alpha matte. The results are shown in Figure 4.

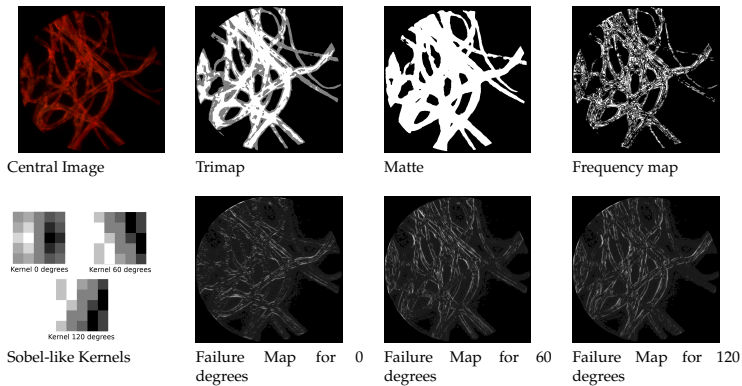


Figure 4. Priors under different form. For coherence, all maps are represented with brighter areas describing higher values. For the failure maps, higher values indicate higher likelihood to fail. In the Sobel-like kernels, bright pixels indicate positive values and dark pixels negative values.

3.2. Cost Volume from Correspondences

A stereo matching is performed to compute a three-dimensional cost volume from the elemental images. The cost function used for this scope is

6. Publications

$$C_{cor}(p) = \alpha_c TAD(p) + (1 - \alpha_c) \xi(p) \quad (5)$$

where we use the notation $C(p)$ to indicate the cost of a pixel p . In this formula the combination of the two terms is controlled by the value $\alpha_c \in [0, 1]$. The first term TAD is the truncated sum of absolute difference and is calculated between two images:

$$TAD(x, y) = \sum_{r=-hws}^{+hws} \sum_{q=-hws}^{+hws} \min(\tau, |I_1(x+r, y+q) - I_2(x+d_x+r, y+d_y+q)|) w_{(x,y,r,q)} \quad (6)$$

In this notation we substitute the pixel p with its x and y coordinates within the image. The sum is then computed on a window, indices r, q are used to reach the pixel within the window and hws indicates half of the size of the window. To deal with three different epipolar lines, the disparity can be seen as a vector $\mathbf{d} = (d, \theta)$ where d is the disparity value and θ the direction ($\theta = 0, 60, 120^\circ$) and is divided into its two horizontal and vertical components, respectively $d_x = d \cos(\theta)$ and $d_y = d \sin(\theta)$. τ is the threshold for truncating the difference.

The second term $\xi(p)$ indicates the census transform and is calculated as:

$$\xi(x, y) = \sum_{r=-hws}^{+hws} \sum_{q=-hws}^{+hws} HD(\xi(I_1(x+r, y+q)), \xi(I_2(x+d_x+r, y+d_y+q))) w_{(x,y,r,q)} \quad (7)$$

where $HD(\cdot)$ indicates the Hamming difference between the census transform of the window around the pixel and the weights are calculated as described in [37]:

$$w_{(x,y,r,q)} = e^{-\left(\frac{|I(x,y) - I(x+r,y+q)|}{\sigma_c} + \frac{dist(x,y,r,q)}{\sigma_d}\right)} \quad (8)$$

where $dist(x, y, r, q) = \sqrt{r^2 + q^2}$ is the euclidean distance between the central pixel and the considered one and σ_c, σ_d are the corresponding parameters regulating the contribution of color and distance.

The cost volume is generated as follows: the correspondences search is completed along the three epipolar lines obtaining $(N_{EI} - 1)$ different cost volumes, where N_{EI} is the number of elemental images ($N_{EI} = 7$ in our case). To merge the cost volumes, the same slice corresponding to a possible disparity value is taken from each cost volume and combined using a weighted average. The failure map calculated in Section 3.1 contains the weights used.

As described in [38], the cost volume is filtered using a guided image filter that takes the colored image as a reference.

3.3. Cost Volume from Defocus

To calculate an accurate defocus map the central image is used as a reference and a difference image for each focal plane is calculated. This allows for a more precise calculation with respect to the defocus response case as defined in [25]: a measure of how much a pixel is in focus at a certain distance.

The difference images are calculated with respect to each focal plane in a similar manner as in the correspondence case:

$$C_{def}(p) = \alpha_d TAD(p) + (1 - \alpha_d) NCC(p) \quad (9)$$

where TAD is defined in Equation (6) and the linear combination of the two terms is controlled by $\alpha_d \in [0, 1]$. The term NCC indicates the normalized cross correlation, calculated as:

$$NCC(x, y) = \sum_{r=-hws}^{+hws} \sum_{q=-hws}^{+hws} \left(\frac{\sigma_{I_1 I_2}^2(x+r, y+q)}{\sigma_{I_1}(x+r, y+q) \sigma_{I_2}(x+r, y+q)} \right) w_{(x,y,r,q)} \quad (10)$$

where the weights $w_{(x,y,r,q)}$ are defined in Equation (8) and $\sigma_{I_1 I_2}, \sigma_{I_1}, \sigma_{I_2}$ indicates respectively the joint variance, the variance of the first and of the second image. The cost volume is built by stacking the

difference images calculated for each focal plane. The cost volume is also filtered using a guided filter as in [38].

3.4. Multi-Scale Approach

Cost volume filtering using a multi-scale approach has shown promising results in refining the cost volume for a higher accuracy of the final depth map [31]. To add consistency and robustness to the proposed cost function, we then adopt a multi-scale approach. Taking inspiration from [31], we build three different layers with a scaling factor $s = 2$, to ensure coherence among images. It has been verified in [31] that building more levels does not significantly improve the final estimation. The cost volume calculated on the smaller scale is then upsampled back and propagated to the initial cost volume.

Differently from [31], however, we do not propagate only the best results of the down-scaled estimation. This idea ensures faster computations, but can lead to larger errors. Instead, we sum the whole cost volume using a weighted average based on the priors information, in this case the frequency map.

$$C_{msc}(p) = w_{s0}(p)C(p) + w_{s1}(p)C_{s1}(p) + w_{s2}(p)C_{s2}(p) \quad (11)$$

As shown in Equation (11), the multi-scale cost (C_{msc}) is a weighted contribution of the cost computed at different scales, being C_{s1} and C_{s2} respectively the costs computed at scale $s1 = 2$ and $s2 = 4$. In Equation (11) $C(p)$ indicates a general cost volume: in our case this is applied to both cost volumes, calculated in Equations (5) and (9). From now on they will be denoted as $C_{def,msc}$ and $C_{cor,msc}$.

The weights w_{si} , $i = 0, 1, 2$ come from the frequency map, that is quantized into a number of levels matching with the number of scales used, and calculated as:

$$w_{si}(p) = \begin{cases} \gamma_1 & \text{if } p \in f_i \\ \gamma_2 & \text{if } p \in f_{i\pm 1} \\ \gamma_3 & \text{otherwise} \end{cases} \quad (12)$$

where we ensure $1 \geq \gamma_1 \geq \gamma_2 \geq \gamma_3 \geq 0$, and typical values are $\gamma_1 = 0.6$, $\gamma_2 = 0.3$ and $\gamma_3 = 0.1$. Here we denote with w_{si} the weight relative to the i -th scale, with γ_{si} the parameter controlling the weight of i -th scale contribution and with f_i and $f_{i\pm 1}$ the i -th and $(i \pm 1)$ -th level of the frequency map.

This allows to shape the cost volume based on the characteristics of the pixels, e.g., pixels that belong to high frequency areas will have a cost based on higher resolution and viceversa, pixel from texture-less regions will have a cost built using the lower resolution.

Tuning the parameters allows us to control the impact of the down-scaled version, and by changing them we can obtain more detail-preserving or smoother depth maps. Note that by setting both $\gamma_1 = \gamma_2 = \gamma_3 = 0.33$ we obtain a standard multi-scale approach that does not make use of the priors information.

3.5. Superpixels

Another technique that reported significant improvement in the cost volume filtering is using superpixels. Superpixels were introduced in [39] and exploited in the depth estimation task [19,32]. The idea behind this consists in grouping pixels with similar characteristics to obtain a more consistent depth estimation. The superpixels are built using two parameters, controlling respectively the approximate size and the similarity between the pixels.

For the depth estimation task there are two main ways of using them: by choosing a small size it can be assumed that the portion of the image corresponding to this superpixel belong to a plane. This allows to compute a single depth value for each superpixel, as done in [19]. This works particularly well for structures that do not exhibit abrupt changes in depth. A different way is shown in [32], where larger size is chosen to allow different depths within a single superpixel. A histogram is built and the best depth estimations are selected and used to filter the cost volume.

6. Publications

Based on these observations we build a flexible approach that takes inspiration from the latter one, but can be restricted to the first. We take the cost volume of the pixels within the superpixels and extract a tentative depth map by extracting the minimum values. From this, a histogram is built, where the highest N_p peaks are selected. Then a penalizing function for the labels not being a peak is built:

$$z(t) = \max \left(0, 1 - \sum_{i=1}^{N_p} \frac{t - \text{ind}(N_i)}{\sigma} \right) \quad (13)$$

where t loops through the labels (different focal planes or disparity values), the maximum function is used to avoid negative functions, $\text{ind}(N_i)$ indicates the index of the i -th peak and σ controls the strictness of the index, i.e., how much a peak is widened to its neighbours. This way of building $z(t)$ ensures $z(t) \in [0, 1]$, that provides an easy handling of the superpixels contribution. In fact, we can sum this back to our cost volume using a penalizing factor. We can then write:

$$C_{sp}(p) = C(p) + \rho c_{sp}(p) \quad (14)$$

where we call C_{sp} the cost volume updated with the superpixel contribution shown as $c_{sp}(p)$, while ρ is the factor that controls the impact of the contribution. In Equation (14) $C(p)$ indicates a general cost volume: in our case this is applied to both $C_{def,msc}$ and $C_{cor,msc}$ to obtain respectively $C_{def,msc,sp}$ and $C_{cor,msc,sp}$. Typical values are $\rho = 0.2$, $N_p = 3$ and $\sigma = 3$.

Note that by changing the parameters and ensuring a smaller size of the superpixels, a maximum number of peak $N_p = 1$ and a small σ , we could obtain a single depth for superpixels, as in [19]. The ρ parameter controls the contribution to the cost volume, with a large value of ρ leading to strongly shape the cost volume for pixels belonging to the same superpixel, and a small value of ρ reducing the cost for having different values within the same superpixel.

3.6. Depth Map Extraction

Many approaches are applicable to extract the depth map from the cost volume. Local and semi-global approaches as winner-takes-all (WTA), semi-global matching (SGM) and more global matching (MGM)[40] are shown to be outperformed from global approaches, where the depth map refinement is posed as an energy minimization problem. The general solution is obtained through minimization of an energy function that depends on two terms, one data term accounting for the cost volume and a smoothness to ensure consistency between neighbouring pixels.

The energy minimization problem can be addressed using several approaches: the most used approaches consist in using Markov random field, as in [25] where the two cost volumes are refined based on their confidence, or in [41] where different energy functions are analyzed, using graph cuts as in [23] or belief propagation as in [42], or a combination of the above methods [43].

We have chosen to use the graph cuts method to minimize an energy function defined as:

$$E(p) = E_{data}(p) + E_{smooth}(p) \quad (15)$$

Intuitively, the data term should include both cost volumes. We used a strategy that allows to cleverly merge the two cubes. By extracting two tentative depth maps using a winner-takes-all approach, some initial considerations can be made. We mainly encounter two cases: one where both guesses agree on a depth value and one where the cost curves have different shapes. The idea here is to apply different weights in these two situations: in the case where there is a large difference, we assume the pixel is unreliable, being either in a texture-less area or part of a repetitive pattern, thus we choose to stick with the defocus estimation, that is most likely to have a guess similar to the real one. In the second case, we choose to give more weight to the correspondence matching, because it is most likely to be more accurate.

We model this by creating an absolute difference map $M_{ad} = \frac{1}{K} |d_{def,wta} - d_{cor,wta}|$, where $d_{def,wta}$ and $d_{cor,wta}$ are respectively the tentative depth map from the defocus and correspondence cost volumes and K is just a normalization factor, that will be used to weight the two contributions of each slice of the cost volume.

Moreover, we want to exploit the points where the two estimations agree. We thus select the reliable pixel where the difference map has its minimum ($M_{ad} = 0$) and for these pixels we penalize the curve with proportion to the distance from the minimum, creating ground control points that are less likely to change their value during the optimization and will serve to distribute the correct value to unreliable neighbours.

The final sum of these contributions can be expressed as

$$E_{data}(p) = (1 - M_{ad}(p))C_{def,msc,sp}(p) + M_{ad}(p)C_{cor,msc,sp}(p) + \beta P_{gcp}(p) \quad (16)$$

where C_{def} and C_{cor} are respectively the cost volume for defocus and correspondences and P_{gcp} is the penalty function used on the ground control points cost curve to enhance the minimum and β is just a scaling factor. The smoothness term ensures consistency across the four adjacent neighbours and can be expressed as $E_{smooth}(p) = \sum_{q \in N} |l(p) - l(q)|$ where N contains the four neighbours of p and $l(p)$ indicates the pixel's label, i.e., its depth value.

The energy is minimized using a multi-linear optimization as explained in [44–46]. The final depth map is filtered using a weighted median filter with integer weights and a guided filter to remove last outliers while preserving structures.

4. Comparative Performance Analysis

The algorithm has been planned for images acquired using the FiMic light field microscope [7], where only visual comparison is possible. Therefore, to prove the worth of the proposed work, additional comparisons were made. We created a set of synthetically generated images with its corresponding ground truth, where numerical analysis can be conducted.

We compared state-of-the-art work on light field microscopy on their available images and state-of-the-art in depth estimation for both real and synthetic images. Methods for light fields are not directly applicable to our images, therefore a fair comparison could not be made.

4.1. State-of-the-Art in Light Field Microscopy

Due to its recent introduction in the microscopy field, there are not many approaches available.

To the best of our knowledge, the best approach in literature belongs to [18], that built a light field microscope using a Lytro first generation camera. The light field in this case consists in 9×9 views of 379×380 pixels, with lateral views exhibiting strong presence of noise. Moreover, the scale of such images, in the order of millimeters, are quite different from the images acquired with FiMic, that can discriminate up to some micrometers.

Despite the image conditions, the extracted depth maps achieve a high accuracy in the reconstruction of the captured objects and maintains consistency in the structure, as shown in Figure 5.

The first image is the head of a daisy, where both approaches reach a satisfying solution, but ours shows a higher level of detail and robustness, visible in the three-dimensional representation.

The second object is an interesting case, where noise, defocus and very low-light condition increases the challenges. Nevertheless, our algorithm is able to reconstruct the structure of the tip of a pencil, that is symmetrical on the y-axis, where the previous approach could not find a solution. This is particularly visible on the lower part of the image.

6. Publications

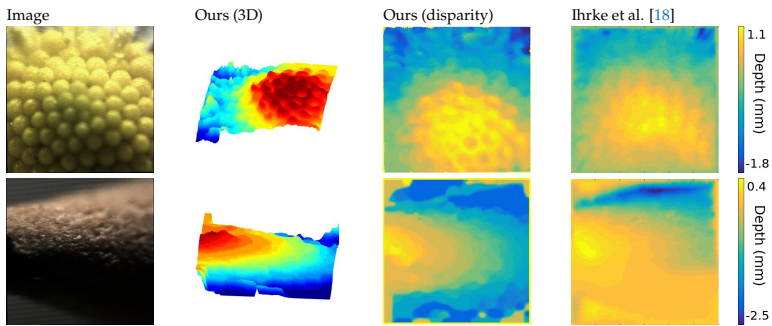


Figure 5. Comparison of depth images from the INRIA dataset. The color coded 3D representation shows the structure of the objects, and the disparity is scaled to try to match the colormap used in [18].

4.2. State-of-the-Art in Depth Map Estimation

To ensure that the proposed approach achieves high quality and accurate results, we also compare against the top performing method for stereo and depth from defocus.

We have chosen to compare with the shape from the focus method, originally published in [47] by taking the best focus measure as analyzed in [48] and a stereo method using neural network to estimate disparity through patches of an image, described in [49]. Unfortunately, we cannot train the network because of lack of dataset, so the pre-trained model is used.

The first set of images, shown in Figure 6, consists of biological samples: the first two images show cotton fibers stained with a fluorescent dye. The third one shows the head of a zebrafish. The target is excited with a single wavelength laser and is emitting light with longer wavelength (i.e., lower energy). With a bandstop filter the laser’s light is filtered out and the light emitted from the sample is captured by the sensor. The wavelength and the color depends on the fluorescent die of the sample.

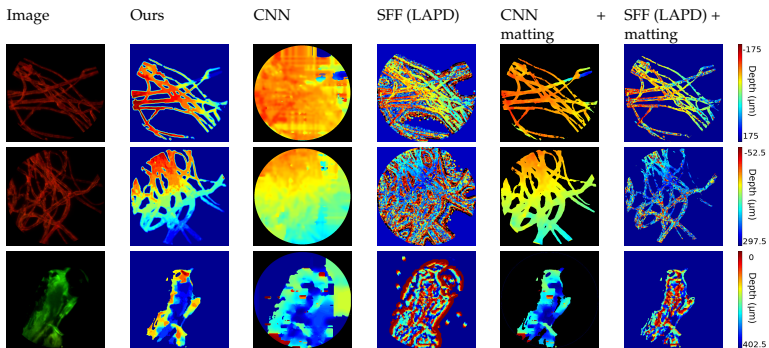


Figure 6. Comparison with Neural Networks (CNN) [49] and Shape from Focus (SFF) [48] on dataset of biological samples: first two rows consist of cotton fibers, last row is the head of a zebrafish.

Therefore, the image exhibits a dark background, where the matting technique described in Section 3.1.3 can be applied. In this case the FiMic was set in order to have: $r = 2.2 \mu\text{m}$, $DOF = 46.4 \mu\text{m}$ and $Z_R = 17.5 \mu\text{m}$.

As expected, approaches from depth map estimation in natural images fail to follow the thin structure because of the lack of texture in the dark region and the presence of repetitive patterns in the fibers.

To ensure a fair comparison, we also applied our matting results to the estimated depth map, obtaining more consistent results. This confirms the high quality of our approach and the robustness of our algorithm that can tackle different kinds of input images.

This being a special case, we also evaluate the results on a different dataset. The second set of images, shown in Figure 7, consists of images of small opaque electrical components. By using a luminance ring it is possible to illuminate the object avoiding most of the shades. For this experiment the FiMic was built differently so it led to: $r = 14.5 \mu\text{m}$, $DOF = 1812 \mu\text{m}$ and $Z_R = 17.5 \mu\text{m}$. Such targets do not need a matting pre-processing step and exhibit structure and texture, being therefore more suitable for standard approaches.

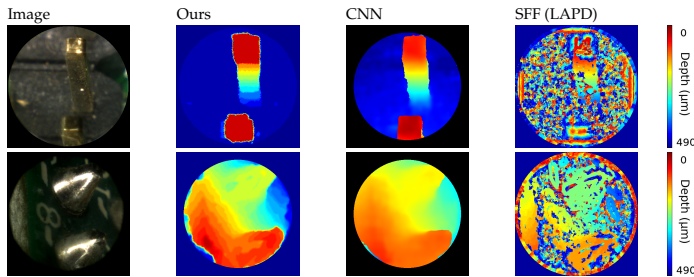


Figure 7. Comparison with CNN [49] and SFF [48] on dataset of opaque electrical components.

Results on these images lead to some considerations: because of the size of the image, the luminance condition and the reflectance of the metal, the depth map calculations are highly challenging, as proven from results of the shape from focus method [48], that recovers only a very noisy reconstruction. Nevertheless, more sophisticated approaches that incorporate filtering steps, as [49], are able to reach satisfying results, obtaining comparable outcomes.

We acknowledge the lack of ground truth images in this field, therefore we propose a numerical evaluation via synthetic images. In recent works Blender has shown to be suitable for emulating light field behaviour and complex scenes [30,50]. We have chosen to use the Blender engine to simulate realistic images of the fibers as if they were captured using the FiMic, thus obtaining the respective ground truth data. To recreate realistic conditions, a special material was generated to emulate the behaviour of semi-transparent fibers and laser illumination. Such images are shown in Figure 8. By doing so, we also obtain the ground truth that allows for a more detailed analysis. Here we show an example of a synthetically generated image and we recap our results in Table 1.

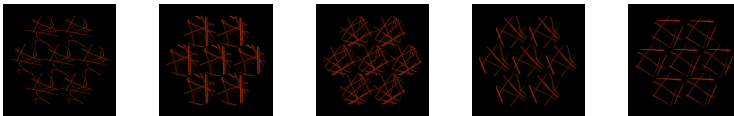


Figure 8. Synthetic images generated with Blender. They simulate the behaviour of the FiMic, shown in Figure 2.

The algorithm was evaluated computing the error as the absolute value of the difference between the estimated disparity and the real one, using the matting mask computed in our approach to reduce

6. Publications

the evaluation to the interesting pixels. Results were averaged over a set of five images with different difficulties and shapes to ensure a fair and consistent comparison. The range of the disparities varies among different images, but is consistent within different approaches.

Table 1. Table with results from synthetic images.

Approach	Error	Standard Deviation
Ours	2.32555	1.8154478
CNN [49]	2.4275436	2.4392762
SFF [48]	9.379839	4.1694072

The error is lower in the proposed approach, as well as the standard deviation, confirming the accuracy and the robustness of the method. The algorithm based on neural network [49] obtains similar performances but still exhibits larger errors, as suggested from the higher variance value, while the shape from focus approach [48] fails in achieving high accuracy as expected, being the most naive approach without post-processing. Nevertheless, it shows that the generated images constitute a challenge for off-the-shelf methods for the disparity estimation and therefore enhance the importance of tackling such challenges.

5. Applications

Depth information can also be used for displaying the 3D information of the microscopic objects. With the technique of [51], one of the views and its corresponding depth map can be used to generate an integral image. This image, projected in an integral imaging (InI) monitor, provides a 3D display of the sample.

To generate the integral image, the view and its corresponding depth map are merged into a 3D point cloud. From this point cloud, a set of synthetic views are computationally generated, and finally processed to obtain the final integral image, which will be projected in the InI monitor. An InI monitor is implemented placing a microlens array in front of a pixelated screen: the lenslets of the MLA integrate the rays proceeding from the pixels to reconstruct the 3D scene. This kind of 3D display is autostereoscopic (glasses-free), it allows multiple observers to experience full parallax and overcomes the accommodation-convergence conflict [52].

We implemented the InI monitor through a Samsung SM-T700 and a MLA composed of lenslets having pitch $p = 1.0$ mm and focal length $f = 3.3$ mm (MLA from Fresneltech, model 630). We used as input for the technique of [51] the RGB and depth images of Figure 7. The integral image obtained and its projection in the 3D InI monitor are shown in Figure 9.

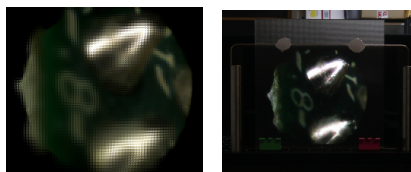


Figure 9. Integral image produced with the method described in [51] and its projection in the 3D monitor.

In this example the image reference plane is set at the background, so the MLA reconstructs the 3D object as a real image, floating in front of the InI monitor. The observer, through his binocular vision system, perceives the depth of the reconstructed objects. A video recording the InI display is visible at the address https://youtu.be/0fPQtckzc_8, in which the parallax and the depth sensation of the reconstructed 3D object is apparent.

In this video, the object shows increasing parallax as we move from far to closer objects. So the left part of the chip and the solder edges move their position with respect to the point of view, while the right part of the chip is almost fixed. This is because during the capture, the chip was tilted with respect to the microscope, as reflected in the depth map of Figure 7.

This technique is very effective for the visualization of the 3D structure of the microscopic samples and proves the usefulness of our work on depth estimation.

6. Summary

The work presented in this paper addresses an important challenge, namely estimating the depth, i.e., the three-dimensional structure of microscopic images. Because of the nature of these images, usually different from natural images, standard approaches may fail. Moreover, due to recent development in the light field technology, light field microscope became a valid alternative for its several applications.

We then use the light field images captured from the recently introduced FiMic [7] that exhibit higher resolution and we showed that the proposed algorithm is capable of accurate reconstruction of challenging scenes, even where previous approaches were failing. These improvements can be helpful for several applications, as presented in the last chapter for the case of lenticular stereoscopic displays, therefore constituting an important contribution for the community.

Appendix A. Implementation Details

Here we provide a table (Table A1) for the parameters relative to the implementation.

Table A1. Parameters Table. The names are consistent with the above formulas which they are referred to. The proposed values are a recommendation and should be adjusted for specific applications. The knowledge of the meaning and the range allow an easier manipulation of such values.

Name	Value	Range	Equation	Meaning
σ_1, σ_2	11, 20	$\sigma_1 > 0, \sigma_2 > 0,$ $\sigma_1 \neq \sigma_2,$	(4)	The variance of the two gaussians kernel used in the difference of gaussians, to build the frequency map.
α_c	0.7	$\alpha_c \in [0, 1]$	(5)	Regulates contributions of TAD and Census in the correspondences cost volume: a higher α leads to higher weights for Census values.
hws	3	$hws > 0,$ $\text{mod}(hws, 2) = 1$	(6), (7), (10)	Half of the size of the windows around the pixel in the cost calculations. Window size can be calculated as $ws = 2 \times hws + 1$.
α_d	0.9	$\alpha_d \in [0, 1]$	(9)	Regulates contributions of TAD and NCC in the defocus cost volume: a higher α leads to higher weights for NCC values.
$\gamma_1, \gamma_2, \gamma_3$	0.6, 0.2, 0.1	$\gamma_1, \gamma_2, \gamma_3 \in [0, 1],$ $\gamma_1 \geq \gamma_2 \geq \gamma_3$	(12)	Used in the multi-scale part. Weights for pixels at different scales with different labels in the frequency map.
σ	3	$\sigma > 0$	(13)	Used in the superpixels section. The width of the peaks in the function penalizing the labels not being a peak.
N_p	3	$N_p > 0$	(13)	Used in the superpixels section. Number of selected peaks. Such peaks will not be penalized. Larger numbers allow more values within the same superpixels.
ρ	0.2	$\rho > 0$	(14)	Controls the weight of the superpixels contribution with respect to the cost volume. $\rho = 0$ cancels superpixels contribution.

The parameters are application dependent, thus the optimal value may not be the same across different executions. We then propose for each parameter a recommended value and the range within which it should vary, followed by a brief explanation of its meaning.

6. Publications

Sensors 2019, xx, 1

14 of 16

Supplementary Materials: Open-source code is available at: <https://github.com/PlenopticToolbox/RobustDepthLFMicroscopy>. Images acquired with the Light Field Microscope (FiMic) are also provided.

Author Contributions: Conceptualization, L.P., G.S. and M.M.-C.; instrumentations, G.S.; methodology, L.P.; applications: N.I.; software, L.P.; resources: L.P., G.S. and N.I.; investigation, L.P.; writing—original draft preparation, L.P. (Sections 1, 3, 4, 6), G.S. (Sections 2) and N.I. (Section 5); writing—review and editing, G.Sa., M.M.-C. and R.K.; funding acquisition, G.Sa., M.M.-C and R.K.

Funding: This project has received funding from the European Union’s Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Actions Grant Agreement No. 676401. On the other hand, M. Martínez-Corral, G. Saavedra and N. Incardona acknowledge funding from Spanish Ministry of Economy and Competitiveness (Grant DPI 2015-66458-C2-1-R).

Acknowledgments: The authors would like to thank Tim Michels (University of Kiel) for the help in preparing the Blender material used to simulate the fibers in the synthetic images.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Levoy, M.; Ng, R.; Adams, A.; Footer, M.; Horowitz, M. Light field microscopy. *ACM Trans. Graph.* **2006**, *25*, 924–934.
2. Levoy, M.; Zhang, Z.; McDowall, I. Recording and controlling the 4D light field in a microscope using microlens arrays. *J. Microsc.* **2009**, *235*, 144–162.
3. Broxton, M.; Grosenick, L.; Yang, S.; Cohen, N.; Andalman, A.; Deisseroth, K.; Levoy, M. Wave optics theory and 3-D deconvolution for the light field microscope. *Opt. Express* **2013**, *21*, 25418–25439.
4. Cohen, N.; Yang, S.; Andalman, A.; Broxton, M.; Grosenick, L.; Deisseroth, K.; Horowitz, M.; Levoy, M. Enhancing the performance of the light field microscope using wavefront coding. *Opt. Express* **2014**, *22*, 24817–24839.
5. Hong, J.Y.; Yeom, J.; Kim, J.; Park, S.G.; Jeong, Y.; Lee, B. Analysis of the pickup and display property of integral floating microscopy. *J. Inf. Disp.* **2015**, *16*, 143–153.
6. Llavador, A.; Sola-Pikabea, J.; Saavedra, G.; Javidi, B.; Martínez-Corral, M. Resolution improvements in integral microscopy with Fourier plane recording. *Opt. Express* **2016**, *24*, 20792–20798.
7. Scrofani, G.; Sola-Pikabea, J.; Llavador, A.; Sanchez-Ortiga, E.; Barreiro, J.; Saavedra, G.; Garcia-Sucerquia, J.; Martínez-Corral, M. FiMic: Design for ultimate 3D-integral microscopy of in-vivo biological samples. *Biomed. Opt. Express* **2018**, *9*, 335–346.
8. Prevedel, R.; Yoon, Y.G.; Hoffmann, M.; Pak, N.; Wetzstein, G.; Kato, S.; Schrödel, T.; Raskar, R.; Zimmer, M.; Boyden, E.S.; et al. Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **2014**, *11*, 727–730.
9. Cong, L.; Wang, Z.; Chai, Y.; Hang, W.; Shang, C.; Yang, W.; Bai, L.; Du, J.; Wang, K.; Wen, Q. Rapid whole brain imaging of neural activity in freely behaving larval zebrafish (*Danio rerio*). *eLife* **2017**, *6*, e28158.
10. Pégard, N.C.; Liu, H.Y.; Antipa, N.; Gerlock, M.; Adesnik, H.; Waller, L. Compressive light-field microscopy for 3D neural activity recording. *Optica* **2016**, *3*, 517–524.
11. Olympus Stream. Industrial Microscope. Available online: <https://www.olympus-ims.com/en/microscope/stream2/#> (accessed on 19.12.2018).
12. Sensofar. 3D Surface Metrology Applications. Available online: <https://www.sensofar.com/metrology/technology/> (accessed on 19.12.2018).
13. Lucideon. Three Dimensional Scanning Electron Microscopy (3D SEM). Available online: <https://www.lucideon.com/testing-characterization/techniques/three-dimensional-scanning-electron-microscopy-3dsem> (accessed on 17.12.2018).
14. Zygo. 3D Optical Surface Profilers. Available online: <https://www.zygo.com/?/met/profilers/> (accessed on 17.12.2018).
15. Filmetrics. The World’s First \$19k 3D Profilometer: The Profilm3D. Available online: <https://www.filmetrics.com/profilometers/profilm3d> (accessed on 19.12.2018).
16. FLIR—Machine Vision. Rolling Down the Cost of 3D Confocal Microscopy. Available online: <https://www.ptgrey.com/case-study/id/10878> (accessed on 18.12.2018).

17. Jung, J.H.; Hong, K.; Park, G.; Chung, I.; Park, J.H.; Lee, B. Reconstruction of three-dimensional occluded object using optical flow and triangular mesh reconstruction in integral imaging. *Opt. Express* **2010**, *18*, 26373–26387. doi:10.1364/OE.18.026373.
18. Mignard-Debise, L.; Ihrke, I. Light-field microscopy with a consumer light-field camera. In Proceedings of the 2015 International Conference on 3D Vision (3DV), Lyon, France, 19–22 October 2015; pp. 335–343.
19. Liu, C.; Narasimhan, S.G.; Dubrawski, A.W. Matting and Depth Recovery of Thin Structures using a Focal Stack. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6970–6978.
20. Wanner, S.; Goldluecke, B. Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 606–619.
21. Zhang, Y.; Lv, H.; Liu, Y.; Wang, H.; Wang, X.; Huang, Q.; Xiang, X.; Dai, Q. Light-field depth estimation via epipolar plane image analysis and locally linear embedding. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 739–747.
22. Chen, C.; Lin, H.; Yu, Z.; Bing Kang, S.; Yu, J. Light field stereo matching using bilateral statistics of surface cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1518–1525.
23. Jeon, H.G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.W.; So Kweon, I. Accurate depth map estimation from a lenslet light field camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1547–1555.
24. Lin, H.; Chen, C.; Kang, S.B.; Yu, J. Depth recovery from light field using focal stack symmetry. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3451–3459.
25. Tao, M.W.; Hadap, S.; Malik, J.; Ramamoorthi, R. Depth from combining defocus and correspondence using light-field cameras. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 673–680.
26. Wang, T.C.; Efros, A.A.; Ramamoorthi, R. Occlusion-aware depth estimation using light-field cameras. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3487–3495.
27. Huang, C.T. Robust Pseudo Random Fields for Light-Field Stereo Matching. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 11–19.
28. Jeon, H.G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.W.; Kweon, I.S. Depth from a Light Field Image with Learning-based Matching Costs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 297–310.
29. Palmieri, L.; Koch, R. Optimizing the Lens Selection Process for Multi-focus Plenoptic Cameras and Numerical Evaluation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1763–1774.
30. Palmieri, L.; Koch, R.; Veld, R.O.H. The Plenoptic 2.0 Toolbox: Benchmarking of Depth Estimation Methods for MLA-Based Focused Plenoptic Cameras. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 649–653.
31. Zhang, K.; Fang, Y.; Min, D.; Sun, L.; Yang, S.; Yan, S.; Tian, Q. Cross-scale cost aggregation for stereo matching. *IEEE Trans. Circ. Syst. Video Technol.* **2017**, *27*, 965–976.
32. Furuta, R.; Ikehata, S.; Yamaskai, T.; Aizawa, K. Efficiency-enhanced cost-volume filtering featuring coarse-to-fine strategy. *Multimed. Tools Appl.* **2017**, *77*, 12469–12491.
33. Lee, Z.; Nguyen, T.Q. Multi-resolution disparity processing and fusion for large high-resolution stereo image. *IEEE Trans. Multimed.* **2015**, *17*, 792–803.
34. Choi, E.; Lee, S.; Hong, H. Hierarchical Stereo Matching in Two-Scale Space for Cyber-Physical System. *Sensors* **2017**, *17*, 1680.
35. Meier, L.; Honegger, D.; Vilhjalmsón, V.; Pollefeys, M. Real-time stereo matching failure prediction and resolution using orthogonal stereo setups. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 5638–5643.
36. Li, C.; Wang, P.; Zhu, X.; Pi, H. Three-layer graph framework with the sumD feature for alpha matting. *Comput. Vis. Image Underst.* **2017**, *162*, 34–45.
37. Yoon, K.J.; Kweon, I.S. Adaptive support-weight approach for correspondence search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 650–656.

6. Publications

38. Hosni, A.; Rhemann, C.; Bleyer, M.; Rother, C.; Gelautz, M. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 504–511.
39. Ren, X.; Malik, J. Learning a classification model for segmentation. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; p. 10.
40. Facciolo, G.; De Franchis, C.; Meinhardt, E. MGM: A Significantly More Global Matching for Stereovision. In Proceedings of the British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; pp. 90–1.
41. Szeliski, R.; Zabih, R.; Scharstein, D.; Veksler, O.; Kolmogorov, V.; Agarwala, A.; Tappen, M.; Rother, C. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1068–1080.
42. Meltzer, T.; Yanover, C.; Weiss, Y. Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In Proceedings of the Tenth IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005; Volume 1, pp. 428–435.
43. Tappen, M.F.; Freeman, W.T. Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In Proceedings of the Eighth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; p. 900.
44. Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1222–1239.
45. Kolmogorov, V.; Zabih, R. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 147–159.
46. Boykov, Y.; Kolmogorov, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1124–1137.
47. Nayar, S.K.; Nakagawa, Y. Shape from Focus; *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 824–831.
48. Pertuz, S.; Puig, D.; Garcia, M.A. Analysis of focus measure operators for shape-from-focus. *Pattern Recognit.* **2013**, *46*, 1415–1432.
49. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **2016**, *17*, 2287–2318.
50. Michels, T.; Petersen, A.; Palmieri, L.; Koch, R. Simulation of plenoptic cameras. In Proceedings of the 2018–3DTV-Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), Helsinki, Finland, 3–5 June 2018; pp. 1–4. doi:10.1109/3DTV.2018.8478432.
51. Incardona, N.; Hong, S.; Martínez-Corral, M.; Saavedra, G. New Method of Microimages Generation for 3D Display. *Sensors* **2018**, *18*, 2805.
52. Son, J.Y.; Javidi, B. Three-dimensional imaging methods based on multiview images. *J. Disp. Technol.* **2005**, *1*, 125–140.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Manuscripts

7.1 Manuscript 1

Geometric Calibration of Multi-Focus Plenoptic Cameras

Nuno Barroso Monteiro, Luca Palmieri, Tim Michels, Leandro Cruz, Reinhard Koch, Nuno Gonçalves and José Gaspar

Submitted to

Journal of Cybernetics, 2020

Reprinted, with permission, from Nuno Barroso Monteiro

The following manuscript is the preprint version submitted to the Journal of Cybernetics. Personal use of this material is permitted. Additional permissions must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Geometric Calibration of Multi-Focus Plenoptic Cameras

Nuno Barroso Monteiro, Luca Palmieri, Tim Michels, Leandro Cruz,
Reinard Koch, Nuno Gonçalves, José António Gaspar

Abstract—A multi-focus plenoptic camera (MPC) images a world point at multiple sensor locations due to the array of microlenses behind the main lens. The multiple focal lengths on the array imply that some points in the images are blurred but, in general, an MPC allows 3D reconstruction from a single image, provided the camera is accurately calibrated. In this work, we propose a new camera model for describing the microlens array of an MPC that considers an unique affine mapping complemented with a blur model associated with each microlens type. The proposed camera model allows to define a calibration procedure using calibration grid corners and their blur radius as features. The features are extracted from the microlens images (MIs) by means of a proposed algorithm that combines local analysis with geometrical region-based refinement. The accuracy of the calibration procedure and the corner detector proposed is evaluated on synthetic and real calibration datasets. The results show that the methods proposed outperform the state of the art.

Index Terms—Blur, Calibration, Corner Detection, Microlens Camera Array, Multi-focus Plenoptic Camera.

I. INTRODUCTION

PLENOPTIC cameras are capable of discriminating the contribution of each light ray emanating from a particular point. The collection of rays captured by these cameras is called a lightfield (LF) [2], [3]. There are several types of plenoptic cameras, namely, standard plenoptic cameras (SPCs) [4], focused plenoptic cameras (FPCs) [5] and, the most recently available on market, multi-focus plenoptic cameras (MPCs) [6].

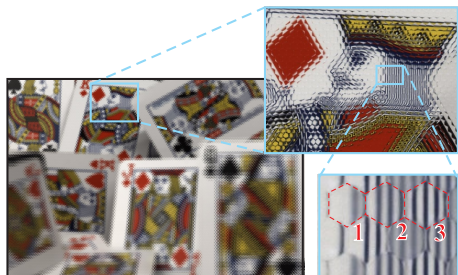
Different plenoptic camera designs gave rise to various, specialized, geometric camera models [7], [8], [9]. Works [10], [11] generalized these models to the different plenoptic cameras but to the best knowledge of the authors, almost no works established relationships between the different camera models. In this work, the different models are studied under a common framework (general model). This general model allows to represent a plenoptic camera despite the different calibration procedures for SPCs, FPCs and MPCs.

This work was supported by the Portuguese Foundation for Science and Technology (FCT) project [grant numbers UID/EEA/50009/2019, and PD/BD/105778/2014] and funded from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

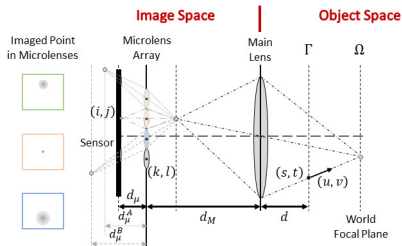
N. B. Monteiro and J. A. Gaspar are with the Institute for Systems and Robotics, University of Lisbon, Portugal, {nmonteiro,jag}@isr.tecnico.ulisboa.pt

L. Palmieri, T. Michels and R. Koch are with the Department of Computer Science, Kiel University, Germany, {lpa,tmr,k}@informatik.uni-kiel.de

L. Cruz and N. Gonçalves are with the University of Coimbra, Institute for Systems and Robotics, and Portuguese Mint and Official Printing Office, INCM Lab, Portugal, lmcruz@isr.uc.pt, nunogon@deec.uc.pt



(a) MPC raw image, zoom of three microlenses



(b) MPC geometry, microlenses with three focal lengths

Fig. 1: Multi-focus effect. (a) Image acquired by an MPC [1]. Small region is augmented to show microlens borders and focusing. MIs, 1 and 2 are blurred, 3 is focused. (b) MPC geometry illustrating the focused and blurred image formation.

Many applications of computer vision like 3D reconstruction require an accurate relationship between the collected LF and the scene. This is significantly dependent on the camera model and the quality of the calibration. In this work, we will focus on MPCs which consist of a main lens, one single high definition imaging sensor, and a microlens array composed of different types of microlenses that differ on their focal plane, *i.e.* focal length. In these cameras, the same scene point is imaged in each microlens type with different degrees of defocus (Figure 1). The geometry of an MPC is based on the geometry of a FPC [5], [6] that generates focused microlens images (MIs) by placing the focal plane of the microlenses on the main lens focal plane.

Strobl *et al.* [12] highlighted the need to model the different microlens types to accurately represent an MPC. Nonetheless,

the only known works that model the different microlens types are [10], [13]. In this work, we define a camera model for describing the microlens array of an MPC based on the work of Monteiro *et al.* [14] considering that the point projections in each microlens are represented by an affine mapping [7] and the defocus present in each microlens type is described by a blur model [15], [16]. The camera model proposed allows to define a calibration procedure for an MPC using corner points and their corresponding blur in each microlens as features. The corner detection in the MIs is particularly challenging in this case, given the different microlens types and defocus blur. Thus, one proposed a detector that separately estimates corner location and radius blur in pixels. The corner location estimation is based on intensity analysis of the boundaries of a window centered around each corner to ensure robustness against different degrees of defocus, while the blur calculation makes use of a conventional focus measure from the literature [17] which is adapted and calibrated for our purposes.

The performance of the corner detector and the calibration procedure proposed is evaluated on synthetic and real calibration datasets. The code and datasets used are provided ¹.

Contributions. The contributions of this work are three-fold: (i) the extension of the affine mapping of Dansereau *et al.* [7] and calibration procedure of Monteiro *et al.* [14] for MPCs, (ii) the definition of the relationships between the affine mapping of Dansereau *et al.* [7] and the camera models in the literature [8], [11] which allow to unify the different models into a single plenoptic camera model, and (iii) the development of a corner detection algorithm that delivers more comprehensive information, *i.e.* position and blur with sub-pixel precision, by combining local and region-based approaches.

In terms of structure, we present in Section II a review of the camera models and calibration methodologies defined for SPCs, FPCs and MPCs. In Section III, we introduce the FPC model considering the camera coordinate system origin at the plane containing the viewpoint projection centers [14]. The mapping from the FPC model to the microlens array representation is described in Section IV. In this section, one also establishes the relationship with the microlens camera model of Bok *et al.* [8]. The extension of the FPC model to the MPC using a model to represent the defocus present in each microlens type is described in Section V. In Section VI, we describe our algorithm to robustly detect, cluster and refine corners in the raw images. The corners are complemented with an estimate of their blur radius that allows to estimate the blur parameters of the MPC. The proposed calibration procedure is described in Section VII with an emphasis in the linear solution based on the microlens array representation. The results of applying the corner detector and the calibration proposed are reported in Section VIII and the major conclusions are presented in Section IX.

Notation: non-italic letters correspond to functions, italic letters correspond to scalars, lower case bold letters correspond to vectors, and upper case bold letters correspond to matrices.

Vectors represented in homogeneous coordinates are denoted by (\cdot) .

II. RELATED WORK

The several works on plenoptic cameras consider the microlenses as pinholes and the main lens as a thin lens regardless of the type of plenoptic camera. One can divide the camera models in the literature in 2D-based and 4D-based mappings. The 2D-based mappings describe the projection of a point in the object space on a particular microlens camera, *i.e.* give the relationship between a point and a pixel. The 4D-based mappings describe the projection of a point onto a collection of 4D rays in the LF.

2D-based Mappings. Johannsen *et al.* [18] and Zeller *et al.* [9] proposed to calibrate an MPC using a single lens type. In these works, the MI center is assumed to lie on the optical axis of the corresponding microlens which causes inaccuracy on the reconstructed points [19]. Additionally, Strobl *et al.* [12] noticed that the calibration of an MPC should consider the different microlens types. Heinze *et al.* [13] used a similar model to [18] accounting for the tilt-shift of the main lens and the different microlens types but not considering an end-to-end image formation. Bok *et al.* [8] performed the calibration of a SPC based on line features extracted from the MIs on the raw image. The model proposed describes a microlens camera using 6 parameters and the knowledge of the corresponding microlens center in the raw image. This SPC calibration method is not robust when the calibration grid is placed near the world focal plane of the main lens since no features are detected on the unfocused MIs [8]. Nousias *et al.* [10] showed that [8] can be extended to a FPC and considered this to calibrate an MPC by performing an independent calibration of each microlens type. Nousias *et al.* [10] acknowledged the existence of common extrinsics among the microlens types but has not proposed a simultaneous calibration of the different microlens types.

In this work, we show that the model proposed by Bok *et al.* [8] can be represented by a 4D-based mapping constraining the microlenses centers coordinates on the raw image to be regularly spaced. This gives further confirmation that a 4D mapping can be extended to model a FPC and MPC.

4D-based Mappings. The mapping of rays defined in pixels (i, j) and microlenses (k, l) indices to rays defined by a position (s, t) and a direction (u, v) in metric units was first proposed by Dansereau *et al.* [7]. This mapping considers a 5×5 matrix with 10 free intrinsic parameters. Monteiro *et al.* [14] represented the 5×5 matrix with 8 free intrinsic parameters by shifting the rays parameterization plane along the optical axis of the camera [20] to the plane containing the viewpoint projection centers and removing the parameters redundant with the extrinsic parameters. Monteiro *et al.* [14] also showed that the 4D mapping can represent a viewpoint camera array which was used to define a methodology for the linear solution step of a calibration based on corner points from viewpoint images (VIs). These models were used to calibrate a virtual SPC that assumes the microlenses define a rectangular tiling on the raw image.

¹URL for dataset and code.

The virtual SPC is obtained after a decoding process [7] to transform the 2D raw image into a 4D LF. This process adds some artifacts [21] that can compromise the quality of the VIs used for the calibration. Zhang *et al.* [11] proposed a generalized model that considers a 5×5 matrix with 6 free intrinsic parameters that is capable of representing the virtual SPC and the FPC. This mapping is identical to the one proposed by Marto *et al.* [22] to describe a camera array composed of cameras with identical intrinsic parameters. Nonetheless, the viewpoint camera array defined by a SPC is not composed of identical cameras since their intrinsics differ on the principal point [14]. In this work, we show that the model proposed by Zhang *et al.* [11] in fact corresponds to a 4D mapping with 8 free intrinsic parameters. There are 2 intrinsic parameters that Zhang *et al.* [11] included in the radial distortion model (Supp. Material F).

The models presented in the literature for the 4D mapping only consider one microlens type. In this work, we complement a 4D based mapping with a blur model for each microlens type to describe an MPC. Additionally, we propose a calibration procedure based on the microlens array representation of the 4D mapping with a nonlinear optimization that minimizes the reprojection and the blur radius errors considering the simultaneous calibration of the microlens types and ensuring common extrinsics among the different microlens types.

In the recent years, several approaches have been proposed for the calibration of plenoptic cameras. However, the most used features in these approaches corresponds to corner points whose detection relies on the generation of VIs [12], [7] where traditional image processing techniques can achieve satisfying results. However, the rendering process for the generation of such VIs for a SPC implies having a calibrated microlens array [10] and, in the case of FPC, having information about the geometry of the scene [6]. Thus, these approaches suffer from a causality dilemma [10].

Nevertheless, different solutions addressing the problem of detecting corners or different features in the raw images have been explored, usually exploiting special physical targets or techniques to recreate favourable conditions for the feature detection. Heinze *et al.* [13] considered a special calibration target with circular pattern to help avoiding incorrect matches along epipolar lines in the depth estimation process. Bok *et al.* [8] claimed that due to the MI small size, corners cannot be accurately detected, and, therefore, edge features of a checkerboard pattern are detected and used for calibration. This approach underperforms in terms of end-to-end image formation and cannot handle different microlens types [10]. Nousias *et al.* [10] operates corners detection on MIs and is able to categorize different microlens types. Corners are found at the saddle point between the two regions of maximum and minimum intensity, ensuring more robustness against blurred MIs. Although outperforming classical state of the art like Harris [23] or FAST [24] corner detectors, it leaves margin for improvement.

In this work, we propose a dedicated corner detector that takes inspiration from and combine ideas from the literature.

Namely, Bok *et al.* [25] analyzed the circular boundaries assuming a sharp change between black and white regions that do not happen for blurred images. Nousias *et al.* [10] used lines towards the highest intensity points to overcome this issue, yet it relies on the image formation process. Our approach does not rely on any image formation process and works on the raw images. We first compute a likelihood map from a boundary centered around a pixel candidate, then remove false matches, divide the corner into clusters and finally apply a refinement step within each cluster to achieve higher accuracy.

III. FOCUSED PLENOPTIC CAMERA

The MPC has a similar geometry to the FPC with a microlens array composed of different types of microlenses. Thus, let us start by defining the camera model for a FPC with a single microlens type.

A FPC can be represented by a 5×5 matrix \mathbf{H} [7], [11] which maps rays $\tilde{\Phi} = [i, j, k, l, 1]^T$ in the image space to rays $\tilde{\Psi} = [s, t, u, v, 1]^T$ in the object (metric) space by

$$\tilde{\Psi} = \mathbf{H} \tilde{\Phi}, \quad (1)$$

where rays $\tilde{\Phi}$ are parameterized using pixels (i, j) and microlenses (k, l) indices and rays $\tilde{\Psi}$ are parameterized using a position (s, t) on a plane Γ and a direction (u, v) defined in metric units [26] (Figure 1.b). The mapping \mathbf{H} defined by Dansereau *et al.* [7] has 12 non-zero entries, however choosing the plane Γ to coincide with the plane containing the viewpoint projection centers and removing the redundancies with the translational components of the extrinsic parameters allows to define the mapping with 8 non-zero entries [14]

$$\mathbf{H} = \begin{bmatrix} h_{si} & 0 & 0 & 0 & 0 \\ 0 & h_{tj} & 0 & 0 & 0 \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (2)$$

This representation is equivalent to the camera model proposed by Zhang *et al.* [11] as demonstrated in the Supp. Material F. In the following we denominate \mathbf{H} as lightfield intrinsics matrix (LFIM)².

IV. MICROLENS CAMERA ARRAY

In this section, we represent a FPC as a camera array of microlenses [8], [27]. The array representation is mapped from the LFIM model (2). Let the projection matrix \mathbf{P}^{kl} , parameterized by the coordinates $(k, l) \in \mathbb{Z}^2$, represent the FPC as an array

$$\mathbf{P}^{kl} = \mathbf{K}^{kl} \left[\mathbf{I}_{3 \times 3} \quad \mathbf{t}^0 + \Delta \mathbf{t}^{kl} \right] \circ \mathbf{T}_w \quad (3)$$

where \mathbf{K}^{kl} denotes the intrinsic matrix, $\mathbf{I}_{3 \times 3}$ is a 3×3 identity matrix, \mathbf{t}^0 denotes the location of the microlens camera

²Notice that LFIM is a simplified term, as \mathbf{H} effectively contains intrinsic parameters information, however, it also contains baseline information, as detailed in Section IV. Conventional extrinsic parameters, as found in pinhole camera models, defining a world coordinate system, are in fact not contained in \mathbf{H} .

array relatively to the camera coordinate system origin, $\Delta \mathbf{t}^{kl}$ corresponds to the location of the microlens projection center relatively to \mathbf{t}^0 , and ${}^c\mathbf{T}_w = \begin{bmatrix} {}^c\mathbf{R}_w & {}^c\mathbf{t}_w \\ \mathbf{0}_{1 \times 3} & \mathbf{1} \end{bmatrix}$ defines the rigid body transformation between the world and camera coordinate systems with rotation ${}^c\mathbf{R}_w \in SO(3)$ and translation ${}^c\mathbf{t}_w \in \mathbb{R}^3$, and $\mathbf{0}_{1 \times 3}$ is the 1×3 null matrix.

Note that while ${}^c\mathbf{T}_w$ defines one coordinate system for all microlens cameras, the intrinsic matrix and the projection center are different for each microlens camera (k, l). In the following, let the camera model for the microlens array (3) take into account that the principal point and the projection center are different for each microlens while the scale factor remains the same:

$$\mathbf{K}^{kl} = \begin{bmatrix} k_u & 0 & u_0 + k \Delta u_0 \\ 0 & k_v & v_0 + l \Delta v_0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

$$\mathbf{t}^0 = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} \text{ and } \Delta \mathbf{t}^{kl} = \begin{bmatrix} k \Delta x_0 \\ l \Delta y_0 \\ 0 \end{bmatrix} \quad (5)$$

where the scalars k_u and k_v denote focal lengths and conversion from metric units to pixels (denominated as scale factors in the remainder of the paper). The vector $[u_0, v_0]^T$ defines the principal point for microlens (k, l) = (0, 0), and the vectors $[\Delta u_0, \Delta v_0]^T$ and $[\Delta x_0, \Delta y_0, 0]^T$ denote principal point shift and baseline between consecutive microlens cameras, respectively. This camera model represents the microlens camera array using 11 parameters.

A. Mapping from LFIM to Microlens Projection Matrices

An arbitrary point $[x, y, z]^T$ along the ray $\Psi = [s, t, u, v]^T$ in the object space can be defined as $[s, t, 0]^T + \lambda[u, v, 1]^T$ for $\lambda \in \mathbb{R}$. Extending the definition of the ray in the object space using the LFIM (2), one obtains the relationship between a 3D point and the ray Φ in the image space [26] as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{H}_{ij}^{st} \begin{bmatrix} i \\ j \end{bmatrix} + z \left(\mathbf{H}_{ij}^{uv} \begin{bmatrix} i \\ j \end{bmatrix} + \mathbf{H}_{kl}^{uv} \begin{bmatrix} k \\ l \end{bmatrix} + \mathbf{h}_{uv} \right) \quad (6)$$

where the LFIM is partitioned in three 2×2 sub-matrices and one 2×1 vector $\mathbf{h}_{uv} = [h_u, h_v]^T$. The sub-matrices follow the notation $\mathbf{H}_{ij}^{(c)}$ where the subscript selects the columns and the superscript selects the lines, i.e. for example, \mathbf{H}_{ij}^{st} selects the first two columns, denoted by ij , and the first two lines, denoted by st .

Considering that the rays of one microlens camera converge to a unique point (s, t) , one may set constant the values (k, l) and solve (6) relatively to (i, j) . This gives an equation of a microlens pixel (i, j) imaging a 3D point (x, y, z) that can be rewritten as a pinhole model, (3) and (5), with the intrinsic matrix and the projection center defined as

$$\mathbf{K}^{kl} = \begin{bmatrix} \frac{1}{h_{ui}} & 0 & -\frac{h_u}{h_{ui}} - k \frac{h_{uk}}{h_{ui}} \\ 0 & \frac{1}{h_{vj}} & -\frac{h_v}{h_{vj}} - l \frac{h_{vl}}{h_{vj}} \\ 0 & 0 & 1 \end{bmatrix}, \quad (7)$$

$$\mathbf{t}^0 = \begin{bmatrix} \frac{h_u}{h_{ui}} h_{si} \\ \frac{h_v}{h_{vj}} h_{tj} \\ \frac{h_{uk}}{h_{ui}} \end{bmatrix} \text{ and } \Delta \mathbf{t}^{kl} = \begin{bmatrix} k \frac{h_{uk}}{h_{ui}} h_{si} \\ l \frac{h_{vl}}{h_{vj}} h_{tj} \\ 0 \end{bmatrix}. \quad (8)$$

This allows to obtain the mappings to the representations in (5). Namely, comparing (8) with (5), we identify a common component $[u_0, v_0]^T = -[h_u/h_{ui}, h_v/h_{vj}]^T$ and a differential (shift) component $[\Delta u_0, \Delta v_0]^T = -[h_{uk}/h_{ui}, h_{vl}/h_{vj}]^T$ for the principal point. The scale factors are defined as $k_u = 1/h_{ui}$ and $k_v = 1/h_{vj}$, and the baseline is defined as $[\Delta x_0, \Delta y_0, 0]^T = [h_{si} h_{uk}/h_{ui}, h_{tj} h_{vl}/h_{vj}, 0]^T$. Finally, the location of the microlens camera array relatively to the camera coordinate system origin is defined as $[x_0, y_0, z_0]^T = [h_{si} h_u/h_{ui}, h_{tj} h_v/h_{vj}, h_{si}/h_{ui}]^T$.

The mapping with the entries of the LFIM allows to redefine the translation vector using the intrinsic parameters in (5) as

$$\mathbf{t}^0 = \begin{bmatrix} u_0 & b_x \\ v_0 & b_y \\ -k_u & b_x \end{bmatrix} \text{ and } \Delta \mathbf{t}^{kl} = \begin{bmatrix} k \Delta u_0 & b_x \\ l \Delta v_0 & b_y \\ 0 & \end{bmatrix}, \quad (9)$$

where $[b_x, b_y, 0]^T = [-h_{si}, -h_{tj}, 0]^T$ corresponds to the baseline between consecutive viewpoint cameras [14]. This allows to represent the microlens camera array with 8 parameters. Furthermore, considering the microlens pinhole constraint (Supp. Material D)

$$k_u b_x = k_v b_y, \quad (10)$$

the microlens camera array can be defined with a minimum of 7 parameters.

B. Mapping from Bok et al. [8] to Microlens Projection Matrices

The model of Bok et al. [8] allows to define a similar projection matrix for the microlens camera (k, l) associated with the microlens center coordinates in the raw image (p_c, g_c) (Supp. Material G) considering

$$\mathbf{K}_b^{kl} = \begin{bmatrix} \frac{f_x}{K_1} & 0 & -\frac{c_x}{K_1} - k \frac{1}{2} \frac{d_h}{K_1} \\ 0 & \frac{f_y}{K_1} & -\frac{c_y}{K_1} - l \frac{d_v}{K_1} \\ 0 & 0 & 1 \end{bmatrix}, \quad (11)$$

$$\mathbf{t}_b^0 = \begin{bmatrix} \frac{K_2}{K_1} \frac{c_x}{K_1} \\ \frac{K_2}{K_1} \frac{c_y}{K_1} \\ \frac{K_2}{K_1} \end{bmatrix} \text{ and } \Delta \mathbf{t}_b^{kl} = \begin{bmatrix} \frac{K_2}{K_1} \frac{1}{2} \frac{d_h}{K_1} \\ \frac{K_2}{K_1} \frac{d_v}{K_1} \\ 0 \end{bmatrix} \quad (12)$$

where K_1 and K_2 are additional intrinsic parameters to the conventional pinhole camera model [28], $c_x = p_0 - c_x$, $c_y = g_0 - c_y$, and the raw image coordinates (p, g) are represented by the 4D coordinates of the rays in the image space using $i = p - p_c$, $j = g - g_c$, and (k, l) using a rectangular sampling basis such that $[p_c, g_c]^T = \text{diag}(d_h/2, d_v) [k, l]^T + [p_0, g_0]^T$.

7. Manuscripts

In this mapping, d_h and d_v correspond to the horizontal and vertical distances between consecutive microlenses centers and (p_0, g_0) correspond to the origin for the (k, l) coordinates in the raw image. The (f_x, f_y, c_x, c_y) are the parameters used to convert normalized coordinates to image coordinates. The camera coordinate system origin corresponds to the plane containing the viewpoint projection centers [8].

Similarly to the microlens intrinsic matrix (8), one can obtain the mapping to the representations in (5). For the principal point, one has $[u_0, v_0]^T = -[\dot{c}_x/K_1, \dot{c}_y/K_1]^T$ and $[\Delta u_0, \Delta v_0]^T = -[d_h/2K_1, d_v/K_1]^T$. The scale factors are defined as $k_u = f_x/K_1$ and $k_v = f_y/K_1$. Finally, the baseline is $[\Delta x_0, \Delta y_0, 0]^T = K_2/K_1 [d_h/2f_x, d_v/f_y, 0]^T$ and the location of the microlens camera array is $[x_0, y_0, z_0]^T = K_2/K_1 [\dot{c}_x/f_x, \dot{c}_y/f_y, 1]^T$. Looking at these definitions, one can identify the same relationships as in (9) with $[b_x, b_y, 0]^T = [-K_2/f_x, -K_2/f_y, 0]^T$.

The representation for the microlens camera (12) allows to identify an incorrect definition for the extrinsic parameters when the z -component of the translation is negative in the calibration procedure proposed by Bok *et al.* [8]. Namely, in this situation, one should change the signs of $r_1, r_2, {}^c\mathbf{t}_w$ and K_2, K_1 is related with the scale factors k_u and k_v in (3) and therefore its sign should not be changed.

V. MULTI-FOCUS PLENOPTIC CAMERA

In the previous sections, we described the camera model of the microlens camera array composed of identical microlenses in a FPC. Nonetheless, for an MPC, one has several types of microlenses each with a different focal length. Considering the thin lens equation to describe a microlens and a fixed distance d_μ , one can see that each microlens type will have a different focal plane. Alternatively, for having a point in the main lens focal plane in focus, each microlens type will need to have a different spacing between the image sensor and the microlens array (d_μ, d_μ^A , and d_μ^B in Figure 1.b). However, there is only one image sensor so some microlenses will produce blurred images of the point. Additionally, the features extracted from the MIs refer to the actual single image sensor at distance d_μ in the MPC and not the virtual image sensors d_μ^A and d_μ^B .

The camera models used for plenoptic cameras consider the microlenses as pinholes [14], [7], [8], [11]. The pinhole model accurately represents the chief-ray originating at a given 3D point. This chief-ray does not depend on the microlens focal plane and detecting its position in the blurred MIs poses a challenge. Thus, in this work, we propose a corner detector that locates the corner point in the blurred MIs as if it was in focus at d_μ (Section VI), and a camera model that describes the point projections of a world point in the different microlenses using a single LFIM (2) and the specific defocus behavior of each microlens type using the blur radius b derived from the models [15], [16]

$$b = s \left| \frac{1}{\mathbf{t}_3 \tilde{\mathbf{m}}} - \frac{1}{z_f} \right| \quad (13)$$

with $s = w d/2$ where w is the distance between the microlens and the image sensor, d is the microlens aperture and z_f is the

depth of the microlens focal plane in the camera coordinate system. $\mathbf{t}_3 \tilde{\mathbf{m}}$ is the depth of the point $\mathbf{m} = [x, y, z]^T$ in the camera coordinate system where \mathbf{t}_3 corresponds to the third row of ${}^c\mathbf{T}_w$. This allows to represent an MPC using an affine mapping with 7 parameters and a blur model with 1 common scale parameter and 1 additional parameter for each microlens type (depth of the microlens focal plane).

VI. CORNERS DETECTION AND EXTRACTION

To be able to accurately calibrate the camera, one relies on the precision of the correspondences detected in the captured images. Our aim is to provide a method capable to detect, cluster and extract corners from images of conventional checkerboards, without need for special calibration targets. The detection is applied directly on the MIs, avoiding dependencies from pre-processing or depth information. We develop a dedicated corner detection algorithm that combines different techniques used in similar works. Our proposed algorithm works in three steps: first we use a boundary approach similar to the one used in [25] to obtain a likelihood map of the corners, then we create clusters of the points belonging to the same checkerboard corners and finally we fit lines to estimate the exact position of the corner within each single cluster.

Moreover, we emphasize the importance of accounting for the different microlens types and thus the degrees of defocus in each MI. The proposed solution is to model this separately from the location, and we achieve this by using a focus measure to estimate the blur radius and incorporate the information in the calibration procedure. So, taking inspiration from a similar idea [29], we generate an ad-hoc dataset of synthetically generated MIs with a corner where we gradually add an increasing amount of blur to simulate all possible blur patterns to reliably map the blur measurements in pixels.

Likelihood Map. The first step of the proposed algorithm consists in the generation of a likelihood map, where each pixel value will indicate the probability of that pixel containing a corner. Such map can be used to extract the actual corners or, as in our case, as an initial estimation step before the refinement step that implies a more sophisticated approach.

We search for corners in the MIs. This allows us to avoid dark areas between MIs and to arbitrarily choose the number of feature per lens, yet it requires knowledge about microlens centers position. This is solved using a white image and the methods described in [7], [8] or using the procedure described in [13], common to Raytrix RxLive software [30] as in [31].

Due to the large number of MI in plenoptic images, each operation will result in a large computational effort. Moreover, we are interested only in the MIs that contain a corner. Based on these considerations, a pre-processing step is performed, classifying the probability of the presence of a corner in each MI by looking at the ratio of dark and light pixels and the presence of lines at different angles.

The calculation of the pixel-wise likelihood map is performed then only on the MIs classified as possible candidates. The proposed method takes inspiration from previous work, where lines towards the highest intensity pixels were drawn

[10] or circular boundaries were analyzed to find the switching point between black and white regions [25].

By merging the two ideas, we calculate a likelihood score by selecting the boundaries of a window around each point and analyzing the curve of the intensity of its values. In our experiment both squared and circular windows were tested, and finally the squared version was chosen. No relevant difference with the circular solution were encountered.

Once we extract the boundaries, we create a linear vector with the values of the pixels intensity. Ideally, such vector should exhibit a particular shape consisting of two distinct maxima and minima value approximately at the same distance between each other, being half of the vector length, as visible in Figure 2.

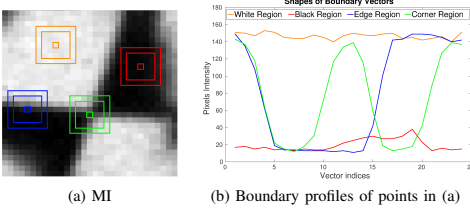


Fig. 2: Example of a MI with a corner, where four different regions relative to white and black textureless areas, an edge and a corner region exhibit their characteristic shapes.

The likelihood score is calculated using two penalty functions that reduce the score when the vector shape differs from the ideal one:

$$L(p) = 1 - \rho_I - \rho_p \quad (14)$$

where $L(p)$ denotes the likelihood score of pixel p and ρ_I , ρ_p are the two penalty functions based respectively on intensity and distance of the peak pixels.

The penalty functions are calculated using the difference between the intensity and relative position values of the minima and maxima and the ideal ones.

We can define the penalty relative to the intensity value as $\rho_I = \sum_{k=1,2} \frac{|I_{m,k} - I_m|}{\sigma_{I,m}} + \frac{|I_{M,k} - I_M|}{\sigma_{I,M}}$ and the one regarding the distance of the peaks as $\rho_p = \frac{|(p_{m,2} - p_{m,1}) - \frac{L}{2}|}{\sigma_{d,m}} + \frac{|(p_{M,2} - p_{M,1}) - \frac{L}{2}|}{\sigma_{d,M}}$ where the m indicates the minima and M the maxima points, I the intensity and p the position in the vector. $\sigma_{d,M}$, $\sigma_{d,m}$, $\sigma_{I,M}$, $\sigma_{I,m}$ are fixed value variable to control the contribution of each penalty function. I_m and I_M are respectively the minimum and maximum intensity value of the whole image, and $\frac{L}{2}$ is half of the vector length.

A final step is required to detect and remove false matches. To avoid assigning scores to pixels that do not actually represent a corner, we run a connected component analysis on a binary version of the map, where all pixels with likelihood greater than zero are selected. If two unconnected components are detected in the same image, we evaluate their score as the

sum of the likelihood of their points and choose the highest one, removing the unwanted matches.

Clustering. The second part of our algorithm consists in clustering. For this a two-dimensional coordinates is required, so we transform our likelihood map into points, by selecting an average position of the pixels that exhibits a likelihood greater than zero. Since we ensured there is only one component per MI, a simple weighted average is enough for our purpose, where the weights used are in fact the likelihood scores.

Before the actual clustering, we filter the points by means of a statistical outliers removal. Outliers are defined as points that do not have enough neighbours within a predefined range. At the same time, we build a grid with a rough guess of where the clusters centers are, to facilitate the convergence of the clustering algorithm. This step is not actually required, yet it significantly reduces the probability of incurring into wrong clustering and the number of iterations needed to reach the final solution.

Following these two steps, we are able to provide as input for the clustering an outlier-free ensemble of points and a rough initial guess of the grid centers. The k-means algorithm has been chosen for the final clustering, using the euclidean distance as measure for the clusters classification.

Refine using Line Constraints. The final part is a region-based process, repeated for every cluster. The operation explained in this section are performed on points within the same cluster. At this step, we do not need two dimensional coordinates, so the likelihood map is again used to achieve higher accuracy in the selection of the final points.

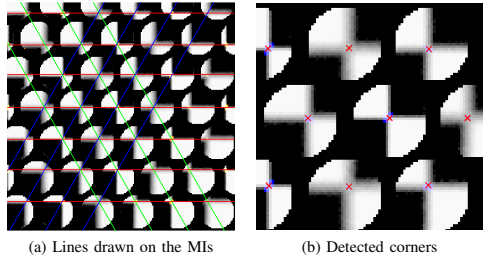


Fig. 3: Example of lines within a single cluster. Each epipolar line is shown in a different color. On the right the likelihood map is shown in a scale of blues and the corners after the refinement with red crosses.

Since we know that the lenses are arranged on a hexagonal grid, assuming rectified images, the epipolar geometry states that the corresponding corners must lie on three epipolar lines, with inclination respectively $[0, 60, 120]$ deg.

Lines can be defined by two parameters, respectively slope and y-intercept. By fixing the slope and tuning the y-intercept value, a score is calculated accumulating the likelihood of the points that lie on each generated line, e.g. a higher score indicates that a line is crossing more high probability points.

Intuitively, the correct lines should be those lines that cross the pixels with higher probability. In order to choose the correct lines and avoid false positives, one must ensure a minimum distance between them to prevent adjacent lines to be chosen together. In our experiment, such distance was set to be the radius of a MI.

The corners lie in the intersection of the generated lines. In an ideal situation, the three lines would intersect in the same point being the corner. In the real case, the distance between the intersections is very small. The final points are chosen if their distance between the intersection with the other lines are smaller than a predefined tolerance threshold.

A larger tolerance allows to select more points at the price of reducing the accuracy, while the choice of a narrow tolerance increases the accuracy but reduces the number of selected points.

Blur Calibration and Estimation. For the case of MPCs, the blur constitutes an important factor that should not be neglected. In our case, the optimal solution consists in incorporating the blur information into our camera model. For that, we estimate the blur radius in pixel for each detected corner. The literature shows different approaches for the measurement of blur. From the analysis conducted in [10] about focus measures on MI for MPCs, we select the Tenengrad Variance method, implemented in [17], which showed promising results.

To incorporate the blur information into the camera model, one need to ensure the consistency of such measure. While in [10] the focus measure was use just to classify the lens type, we aim at precisely estimating the blur radius in pixels.

Our proposed solution is based on: i) estimating locally in a smaller region around the detected corner instead of using the focus measure on the whole MI. The remaining part of the image should not affect our estimation, since corners at the edge of a microlens may have less texture and thus obtain different focus measurements. ii) creating a dedicated set of images with a fixed size equal to the region used, where we gradually increase the blur to obtain a series of templates of corners with different amount of blur, ensuring the consistency relation between blur and the focus measure,

In terms of implementation, we create a lookup-table for the relationship between focus measure and blur radius in pixels. Then, for each corner we detect in a MI, we apply the focus operator and we fetch the desired blur value.

VII. CALIBRATION

The proposed calibration is based on finding the corners of a planar calibration grid of known dimensions and the corresponding blur radius as features. In the following, we assume that the microlens centers and types are known [10], [7] and that the corners in the world coordinate system have been matched with the imaged corners. An imaged corner is defined by a ray $\Phi = [i, j, k, l]^T$ in the image space. The (i, j) coordinates correspond to the pixel coordinates of the detected corners on the MIs relatively to the corresponding microlens center. The (k, l) coordinates correspond to the microlens coordinates considering a rectangular sampling to represent the microlens center coordinates in the raw image (Figure 4.b).

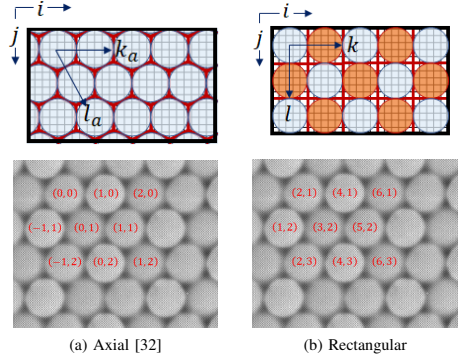


Fig. 4: Axial (a) and rectangular (b) coordinate systems to represent the microlens centers in the raw image.

A. Linear Initialization

In this section, we will consider the mapping in Section IV to define a linear solution for the microlens array (3) associated with a plenoptic camera and the extrinsic parameters for each pose of the calibration grid. The blur model (13) described in Section V associated with each microlens type is used to define the microlens focal planes.

Homography Estimation. Considering the microlens projection matrix (3), a point $\mathbf{m} = [x, y, z]^T$ in the object space is projected to a point in the image plane \mathbf{q} by

$$\tilde{\mathbf{q}} \sim \mathbf{P}^{kl} \tilde{\mathbf{m}} = \mathbf{K}^{kl} \begin{bmatrix} \mathbf{c}\mathbf{R}_w & \mathbf{c}\mathbf{t}_w + \mathbf{t}^{kl} \end{bmatrix} \tilde{\mathbf{m}} \quad (15)$$

where the symbol \sim denotes equal up to a scale factor and $\mathbf{t}^{kl} = \mathbf{t}^0 + \Delta \mathbf{t}^{kl}$. The co-planar grid points allow to define a world coordinate system such that the z -coordinate is zero. In this context, denoting $\tilde{\mathbf{m}} = [x, y, 1]^T$, one can redefine the projection (15) as $\tilde{\mathbf{q}} \sim \mathbf{H}^{kl} \tilde{\mathbf{m}}$ where

$$\mathbf{H}^{kl} = \mathbf{K}^{kl} \begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{c}\mathbf{t}_w + \mathbf{t}^{kl} \end{bmatrix} \quad (16)$$

is the parametric homography matrix for the microlens camera (k, l) , and $\mathbf{c}\mathbf{R}_w = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$. The homography matrix \mathbf{H}^{kl} like the projection matrix (3) changes among microlenses as a result of the principal point shift and baseline in (8).

Let us consider that \mathbf{H}^{kl} can be defined from the homography matrix \mathbf{H}^0 associated with the microlens coordinates $(k, l) = (0, 0)$ and the homography microlens change matrix \mathbf{A}^{kl} by

$$\mathbf{H}^{kl} = \underbrace{\begin{bmatrix} h_{11}^0 & h_{12}^0 & h_{13}^0 \\ h_{21}^0 & h_{22}^0 & h_{23}^0 \\ h_{31}^0 & h_{32}^0 & h_{33}^0 \end{bmatrix}}_{\mathbf{H}^0} + \begin{bmatrix} k & 0 & 0 \\ 0 & l & 0 \\ 0 & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 0 \end{bmatrix}}_{\mathbf{A}^{kl}}. \quad (17)$$

Considering the homography projection of a calibration grid corner $\tilde{\mathbf{m}} = [x, y, 1]^T$ in the object space to the image

point $\tilde{\mathbf{q}}$ for the microlens camera (k, l) , applying the cross product by $\tilde{\mathbf{q}}$ on each side of the projection equation leads to $[\tilde{\mathbf{q}}]_{\times} \mathbf{H}^{kl} \tilde{\mathbf{m}} = \mathbf{0}_{3 \times 1}$, where $[(\cdot)]_{\times}$ is a skew-symmetric matrix that applies the cross product. Using the properties of the Kronecker product [33] and solving for each of the unknown parameters, one obtains

$$\left(\tilde{\mathbf{m}}^T \otimes [\tilde{\mathbf{q}}]_{\times} \right) \mathbf{T} \begin{bmatrix} \mathbf{h}^0 \\ \mathbf{a}^{kl} \end{bmatrix} = \mathbf{0}_{3 \times 1} \quad (18)$$

where

$$\mathbf{T} = \begin{bmatrix} k & 0 & 0 & 0 & 0 & 0 \\ 0 & l & 0 & 0 & 0 & 0 \\ & & \mathbf{0}_{1 \times 6} & & & \\ \mathbf{I}_{9 \times 9} & 0 & 0 & k & 0 & 0 \\ 0 & 0 & 0 & 0 & l & 0 \\ & & \mathbf{0}_{1 \times 6} & & & \\ 0 & 0 & 0 & 0 & k & 0 \\ 0 & 0 & 0 & 0 & 0 & l \\ & & \mathbf{0}_{1 \times 6} & & & \end{bmatrix}, \quad (19)$$

and \mathbf{h}^0 and \mathbf{a}^{kl} correspond to vectorizations of the matrix \mathbf{H}^0 and \mathbf{A}^{kl} by stacking their columns and removing the zero entries, respectively. The solution $[\mathbf{h}^0, \mathbf{a}^{kl}]^T$ for the parametric homography matrix can be estimated using singular value decomposition (SVD).

The parametric homography matrix (17) is defined using 15 parameters. In an MPC, each point in the object space originates multiple image points and according to equation (18), each point correspondence $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$ originates three equations with only two being linearly independent. Nonetheless, the restrictions on the microlens camera array also originate restrictions on the projections of a point in the object space. Namely, the ray in the image space $\Phi^{kl} = [i, j, k, l]^T$ associated with an arbitrary microlens (k, l) can be described from the ray coordinates $\Phi^0 = [i_0, j_0, 0, 0]^T$ associated with the microlens $(k, l) = (0, 0)$ by $\Phi^{kl} = \Phi^0 + [k\beta, l\beta, k, l]^T$, where β corresponds to the disparity of the point defined on the MIs. This reduces the number of linearly independent equations originated by a point in the object space to four [14]. Thus, one needs at least four non-collinear points to obtain the entries of the homography matrix \mathbf{H}^{kl} .

Intrinsic and Extrinsic Estimation. The structure of the homography matrix (16) in conjunction with the orthogonality and identity of the column vectors of ${}^c\mathbf{R}_{uv}$ allow to define constraints on the intrinsic parameters as $\mathbf{h}_1^T \mathbf{B}^{kl} \mathbf{h}_2 = 0$ and $\mathbf{h}_1^T \mathbf{B}^{kl} \mathbf{h}_1 - \mathbf{h}_2^T \mathbf{B}^{kl} \mathbf{h}_2 = 0$ [34] where \mathbf{h}_n refers to the n -th column vector of \mathbf{H}^{kl} , and the symmetric matrix that describes the image of the absolute conic is defined as $\mathbf{B}^{kl} = \mathbf{K}^{kl-T} \mathbf{K}^{kl-1}$ [34], [35]. Using the knowledge of the intrinsic matrix defined in Section IV-A, one can represent the absolute conic \mathbf{B}^{kl} for a microlens camera (k, l) using a minimal number of parameters.

The intrinsic matrix \mathbf{K}^{kl} differs on the principal point for each microlens leading to different images of the absolute conic. The principal points change regularly between consecutive microlenses by $\begin{bmatrix} -\frac{h_{uk}}{h_{ui}} & -\frac{h_{vl}}{h_{vj}} \end{bmatrix}^T$ which can be used

to constraint the parametric representation of \mathbf{B}^{kl} . Namely, considering (8), \mathbf{B}^{kl} can be defined as

$$\mathbf{B}^{kl} = \mathbf{B}^0 + k \mathbf{C}^k + l \mathbf{D}^l + k^2 \mathbf{E}^k + l^2 \mathbf{F}^l \quad (20)$$

with

$$\mathbf{B}^0 = \begin{bmatrix} h_{ui}^2 & 0 & h_u h_{ui} \\ 0 & h_{vj}^2 & h_v h_{vj} \\ h_u h_{ui} & h_v h_{vj} & 1 + h_u^2 + h_v^2 \end{bmatrix}, \quad (21)$$

$$\mathbf{C}^k = \begin{bmatrix} 0 & h_{ui} h_{uk} & 0 \\ 0 & 0 & 0 \\ h_{ui} h_{uk} & 0 & 2h_u h_{uk} \end{bmatrix}, \quad \mathbf{E}^k = \begin{bmatrix} 0 & \mathbf{0}_{2 \times 3} \\ 0 & h_{uk}^2 \end{bmatrix}, \quad (22)$$

$$\mathbf{D}^l = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & h_{vj} h_{vl} \\ 0 & h_{vj} h_{vl} & 2h_v h_{vl} \end{bmatrix}, \quad \text{and } \mathbf{F}^l = \begin{bmatrix} 0 & \mathbf{0}_{2 \times 3} \\ 0 & h_{vl}^2 \end{bmatrix}. \quad (23)$$

This allows to define a representation for \mathbf{B}^{kl} using 11 distinct non-zero entries $\mathbf{b}^{kl} = [b_{11}, b_{13}, b_{22}, b_{23}, b_{33}, c_{13}, c_{33}, d_{23}, d_{33}, e_{33}, f_{33}]^T$ where $(\cdot)_{mn}$ represents the entry in row m and column n of the matrix (\cdot) . Considering these parameters, the intrinsic parameters constraints can be redefined as

$$\begin{bmatrix} h_{11} h_{12} & h_{11}^2 - h_{12}^2 \\ h_{11} h_{32} + h_{12} h_{31} & 2(h_{11} h_{31} - h_{12} h_{32}) \\ h_{21} h_{22} & h_{21}^2 - h_{22}^2 \\ h_{21} h_{32} + h_{22} h_{31} & 2(h_{21} h_{31} - h_{22} h_{32}) \\ h_{31} h_{32} & h_{31}^2 - h_{32}^2 \\ k(h_{11} h_{32} + h_{12} h_{31}) & 2k(h_{11} h_{31} - h_{12} h_{32}) \\ k(h_{31} h_{32}) & k(h_{31}^2 - h_{32}^2) \\ l(h_{21} h_{32} + h_{22} h_{31}) & 2l(h_{21} h_{31} - h_{22} h_{32}) \\ l(h_{31} h_{32}) & l(h_{31}^2 - h_{32}^2) \\ k^2(h_{31} h_{32}) & k^2(h_{31}^2 - h_{32}^2) \\ l^2(h_{31} h_{32}) & l^2(h_{31}^2 - h_{32}^2) \end{bmatrix}^T \mathbf{b}^{kl} = \mathbf{0}_{2 \times 1}. \quad (24)$$

Normally, each homography generates two equations for determining the matrix of the absolute conic image [34]. The parametric representation (17), representing an arbitrary microlens (k, l) , generates six equations. Nonetheless, only two equations are independent regarding the entries of \mathbf{B}^0 , so one needs to acquire at least three calibration grid poses to estimate \mathbf{b}^{kl} defined up to a scale factor.

The intrinsic matrix parameters can be recovered from \mathbf{B}^{kl} . More specifically, rewriting the intrinsic matrix \mathbf{K}^{kl} (5) as

$$\mathbf{K}^{kl} = \underbrace{\begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K}^0} + \begin{bmatrix} k & 0 & 0 \\ 0 & l & 0 \\ 0 & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} \Delta u_0 & \\ & \Delta v_0 \\ & & 0 \end{bmatrix}}_{\mathbf{C}^{kl}}, \quad (25)$$

one can define $\mathbf{B}^0 = \mathbf{K}^{0-T} \mathbf{K}^{0-1}$. This allows to estimate the entries of \mathbf{K}^0 using the Cholesky decomposition of \mathbf{B}^0 and correcting the scale factor considering $k_{33}^0 = 1$. The principal point shift can be estimated considering $\Delta u_0 = -\frac{h_{uk}}{h_{ui}} = -\frac{c_{13}}{b_{11}}$ and $\Delta v_0 = -\frac{h_{vl}}{h_{vj}} = -\frac{d_{23}}{b_{22}}$.

7. Manuscripts

9

The extrinsic parameters can be estimated once the intrinsic matrix \mathbf{K}^{kl} is known. From (16), the rotation matrix ${}^c\mathbf{R}_w = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$ is recovered considering

$$\mathbf{r}_1 = \lambda \mathbf{K}^{kl^{-1}} \mathbf{h}_1, \quad \mathbf{r}_2 = \lambda \mathbf{K}^{kl^{-1}} \mathbf{h}_2, \quad \text{and} \quad \mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2 \quad (26)$$

with $\lambda = 1 / \|\mathbf{K}^{kl^{-1}} \mathbf{h}_1\| = 1 / \|\mathbf{K}^{kl^{-1}} \mathbf{h}_2\|$. The translation ${}^c\mathbf{t}_w$ and projection center \mathbf{t}^{kl} (9), considering the microlens pinhole restriction (10), are recovered solving the following system of equations

$$\lambda \mathbf{h}_3 = \begin{bmatrix} \mathbf{K}^{kl} & \mathbf{K}^{kl} \mathbf{J} \end{bmatrix} \begin{bmatrix} {}^c\mathbf{t}_w \\ b_x \end{bmatrix} \quad (27)$$

with

$$\mathbf{J} = \begin{bmatrix} u_0 + k\Delta u_0 \\ (v_0 + l\Delta v_0) \frac{k_x}{k_v} \\ -k_u \end{bmatrix}. \quad (28)$$

Blur Estimation. The blur model described in Section V defines the defocus that occurs in each microlens depending on the distance of the point to the microlens focal plane. The depth of the microlens focal plane corresponds to the depth of the points with blur radius equal to zero (13), *i.e.* $z_f = \mathbf{t}_3 \mathbf{m}$ for $b = 0$. Normally, the blur radius is not zero due to pixel discretization so one should consider a range for selecting the points with zero blur radius and take the median of the points depth to estimate the microlens focal plane. Once the microlens focal plane depth is known, the parameter s is estimated simply taking the median of $s = \frac{1}{b} \left| \frac{1}{\mathbf{t}_3 \mathbf{m}} - \frac{1}{z_f} \right|$.

B. Nonlinear Optimization

In this section, the linear solution is refined and radial distortion is considered on the coordinates (u, v) . Namely, the undistorted rays in the object space $\Psi^u = [s, t, u^u, v^u]^T$ are defined from distorted rays in the object space $\Psi = [s, t, u, v]^T$ by

$$\begin{bmatrix} u^u \\ v^u \end{bmatrix} = \left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 \right) \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} + \begin{bmatrix} b_u \\ b_v \end{bmatrix} \quad (29)$$

where $\hat{u} = u^u - b_u$, $\hat{v} = v^u - b_v$, $r^2 = u^2 + v^2$, and $\mathbf{d} = (k_1, k_2, k_3, b_u, b_v)$ defines the distortion vector. In the distortion vector, k_1 , k_2 and k_3 are the radial distortion correction coefficients while the vector $[b_u, b_v]^T$ defines the distortion center. In the nonlinear optimization, we minimize the reprojection error $\Theta(\cdot)$ and the blur radius error $\tau(\cdot)$ simultaneously for all microlens types

$$\arg \min_{\mathbf{K}^{kl}, \mathbf{t}^{kl}, \mathbf{b}_m, \mathbf{R}_p, \mathbf{t}_p, \mathbf{d}} \Theta(\mathbf{K}^{kl}, \mathbf{t}^{kl}, \mathbf{d}, \mathbf{R}_p, \mathbf{t}_p) + \tau(\mathbf{b}_m, \mathbf{R}_p, \mathbf{t}_p). \quad (30)$$

This optimization refines the intrinsic parameters \mathbf{K}^{kl} and \mathbf{t}^{kl} , the blur parameters $\mathbf{b}_m = [s, z_m]^T$, $m = 1, \dots, M$ where M is the number of microlens types, the extrinsic parameters \mathbf{R}_p (parameterized by Rodrigues formula [28]) and \mathbf{t}_p , $p =$

$1, \dots, P$ where P is the number of poses, and the distortion vector \mathbf{d} .

The reprojection error [36]

$$\Theta(\mathbf{K}^{kl}, \mathbf{t}^{kl}, \mathbf{d}, \mathbf{R}_p, \mathbf{t}_p) = \sum_{p=1}^P \sum_{n=1}^{N_p} \sum_{(k,l) \in \chi_n} \left\| \hat{\mathbf{q}}_n^{kl} - \mathbf{q}_n^{kl} \right\|^2 \quad (31)$$

defines the error in pixels using the Euclidean distance between the detected corners $\hat{\mathbf{q}}_n^{kl}$ and the projections \mathbf{q}_n^{kl} of the world coordinate system point \mathbf{m}_n associated with the corner n in the multiple microlens cameras χ_n , *i.e.* $\mathbf{q}_n^{kl} = \Pi(\mathbf{R}_p \mathbf{m}_n + \mathbf{t}_p)$ where $\Pi(\cdot)$ defines the multiple projections of a point in the camera coordinate system. The detected corners are not directly the ones obtained from the raw image but the projections obtained from the reconstructed point after distortion correction, *i.e.* $\hat{\mathbf{q}}_n^{kl} = \Pi(\eta(\mathbf{H}_n \mathbf{d}, \Phi_n))$ where η defines the reconstructed point after mapping the ray in the image space Φ_n associated with the corner n to the ray in object space (1), followed by distortion rectification (29) and reconstruction [26]. N_p corresponds to the number of corners detected on a pose p .

The blur radius

$$\tau(\mathbf{b}_m, \mathbf{R}_p, \mathbf{t}_p) = \sum_{p=1}^P \sum_{m=1}^M \sum_{n=1}^{N_p} \sum_{(k,l) \in \chi_n} \left\| \hat{b}_n^{kl} - b_n^{kl} \right\|^2 \quad (32)$$

defines the error in pixels using the Euclidean distance between the detected blur radius \hat{b}_n^{kl} and the blur radius b_n^{kl} estimated for the point \mathbf{m}_n using (13) for the multiple microlens cameras.

The nonlinear optimization is solved using the trust-region-reflective algorithm [37], where a sparsity pattern for the Jacobian matrix is provided. The number of parameters over which we optimize is 7 for the intrinsic parameters, $M + 1$ for the blur parameters, 5 for the lens distortion parameters, and $6P$ for the extrinsic parameters.

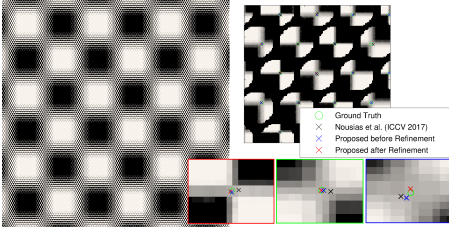
VIII. RESULTS

In this section, one will present the results of the corner detector and the calibration procedure proposed in Sections VI and VII, respectively. The methodologies proposed are applied to synthetic datasets obtained using the toolbox [38] and to a dataset acquired with a commercially available MPC with 3 microlens types, the R42 Raytrix with a 50 mm lens.

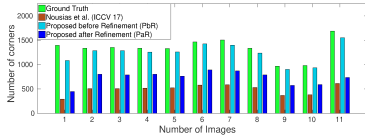
A. Synthetic Datasets and Corners Detection

Synthetic images have successfully proven to emulate plenoptic cameras [38], so one created a dedicated set of synthetic raw images with a checkerboard pattern using the Blender engine.

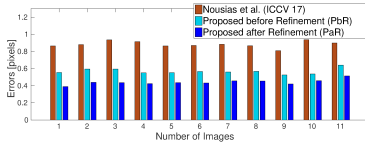
The knowledge of the three dimensional position of the pattern and the camera parameters allows to calculate the corners' positions on the rendered image using ray tracing. For every pixel, a bundle of rays is emitted and traced to the scene until they reach the object, fetching its position in the three-dimensional space. Even in the unfocused case, in which the



(a) Synthetic image and detected corners.



(b) Number of detected corners per image.



(c) Ground truth based average corner location error.

Fig. 5: Results relative to the synthetic dataset. (a) shows a checkerboard image and examples detected corners for different microlens types and different amounts of blur, (b) relates to the number of corners found and (c) reports the average error in pixels.

rays may not converge to the same point, their positions can be averaged to robustly recover the pixel’s positional information [39]. One way to do this is to render a positional image along with the colored image and matching the colored point with its positional information. By rendering the positional image with a higher resolution and picking the closest point, we can reach a sub-pixel accuracy close to 0.1 pixels. This information allows us to create a benchmark and evaluate the performance of the corner detection algorithm proposed.

For this purpose, a set of 11 images of 3500×3500 pixels and 11017 MIs is created. On average, each image contains 1335.4 corners, for a total of 14689 corners, ensuring the statistical significance of our analysis. Standard corner detection methods have shown to fail on MIs, so we performed a performance comparison analysis only against the state of the art [10] (denoted as *Nousias17*). Additionally, we show the proposed method performance before and after the refinement stage described in Section VI to give a further insight on how the proposed method works. Errors are calculated as the difference in pixels between the estimated corner point position and the corresponding ground truth point. In Figure 5.c, the average error for the detected corners in each synthetic

image is shown. Our method retains a lower error in all the images, and it is possible to see how the refinement step improves the estimation, reducing the errors in average by 22.05%.

For a meaningful analysis of the error, the number of detected corners has to be taken into account. In Figure 5.b, we reported the number of corners detected from each algorithm alongside with the number of ground truth corners for each synthetic image.

TABLE I: Summary of the average results per pose obtained using the different corner detectors. The highlighted values indicates the best result for each category.

Average Results	Error [pix]	Corners Detected	Detection Ratio
<i>Nousias17</i> [10]	0.8842	491.09	36.88%
Proposed	0.5677	1237.8	92.79%
Proposed Refined	0.4420	731.55	55.29%

In Table I, we summarize the corner detection results indicating the average error between the estimated and ground truth corners and the average number of corners detected per pose. The detection ratio gives the percentage of corners detected with respect to the actual number of corners imaged in the synthetic dataset. The proposed method outperforms the state of the art [10]. As expected, the initial step before refinement aims at detecting all corners, reaching almost the full score. The quantity is then traded with the quality in the refinement step.

B. MPC Calibration Results

MPCs have a microlens array composed of differently focused microlenses. For example, the Raytrix camera has three types of microlenses that exhibit different degrees of defocus (Figure 1.a). Thus, the camera model proposed in Section V considers a blur model for each microlens type and the calibration procedure considers the detected corner points and blur radius as features (Figure 6).

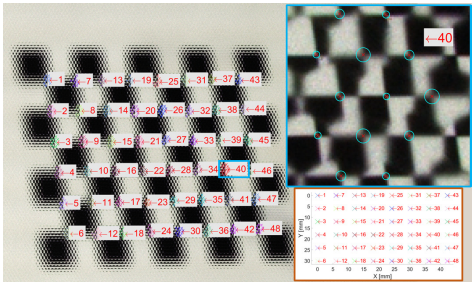


Fig. 6: Dataset acquired with a Raytrix camera. The corners and blur radius detected on the raw image and the clustering performed by the algorithm proposed are highlighted. A detail of cluster 40 is shown in the blue rectangle. The association of the clusters with the 3D points is depicted in the orange rectangle.

TABLE II: RMS reprojection error in pixels for synthetic dataset considering different calibration procedures and corner detectors. The highlighted values correspond to the best result for a given stage of the calibration. * denotes the calibration procedure defined by Nousias *et al.* [10] with the correction in Section IV-B.

Reprojection Error [pix]		Corner Detector		
Calibration Procedure		Ground Truth	<i>Nousias17</i> [10]	Proposed
Initial	<i>Nousias17</i> [10]	7.939	4.684	7.887
	<i>Nousias17*</i> [10]	17.155	3.541	12.402
	Proposed	0.393	2.249	0.950
Optimized	<i>Nousias17</i> [10]	0.720	1.202	4.273
	<i>Nousias17*</i> [10]	0.197	0.727	0.534
	Proposed	0.216	0.748	0.535
Optimized (with Distortion)	Proposed	0.213	0.743	0.528

TABLE III: RMS reconstruction error in mm for synthetic dataset considering different calibration procedures and corner detectors. The highlighted values correspond to the best result for a given stage of the calibration. * denotes the calibration procedure defined by Nousias *et al.* [10] with the correction in Section IV-B.

Reconstruction Error [mm]		Corner Detector		
Calibration Procedure		Ground Truth	<i>Nousias17</i> [10]	Proposed
Initial	<i>Nousias17</i> [10]	3154.9	1795.6	775.9
	<i>Nousias17*</i> [10]	133.4	86.8	13.9
	Proposed	2.3	18.9	5.0
Optimized	<i>Nousias17</i> [10]	52.3	55.6	485.5
	<i>Nousias17*</i> [10]	39.7	53.9	8.3
	Proposed	1.1	6.6	5.3
Optimized (with Distortion)	Proposed	1.4	10.6	5.7

Synthetic Dataset Results. In this section, the performance of the corner detectors is evaluated on the estimation of the synthetic MPC parameters considering three different sets of corners: (i) the ground truth corners provided by the synthetic dataset, and (ii) the corners detected by the algorithm proposed by Nousias *et al.* [10] and (iii) by the proposed algorithm (Section VI). These corners are used by the proposed calibration procedure (Section VII) and the state of the art calibration procedure for MPCs [10] (denoted as *Nousias17*). For this comparison, we consider the root mean square (RMS) of the reprojection and reconstruction errors for the different stages of the calibration process: the initial linear solution and the nonlinear refinement with and without distortion estimation. The results obtained are summarized in Tables II and III.

In Tables II and III, the reprojection and reconstruction errors for the calibration proposed using the ground truth corners attain small values which shows that the camera model defined in Section V is suitable to represent MPCs. The reprojection error is similar to the one obtained using the correction defined in Section IV-B for the state of the art calibration of Nousias *et al.* [10] while the reconstruction error obtained using the calibration proposed is significantly smaller. One should highlight that the proposed camera model does not need to know the position of the microlenses centers, contrarily to the method of Nousias *et al.* [10].

The correction proposed in Section IV-B to the calibration procedure of Nousias *et al.* [10] provides better results than applying directly the methodology of Nousias *et al.* [10]. Namely, the reprojection error decreases by 72.6% and the

TABLE IV: RMS reprojection and reconstruction errors for Raytrix dataset considering different calibration procedures and corner detectors. The highlighted values correspond to the best result for a given stage of the calibration. * denotes the calibration procedure defined by Nousias *et al.* [10] with the correction in Section IV-B.

Calibration Procedure		Corner Detector			
		Reprojection Error [pix]		Reconstruction Error [mm]	
		<i>Nousias17</i> [10]	Proposed	<i>Nousias17</i> [10]	Proposed
Initial	<i>Nousias17</i> [10]	9.437	4.899	2158.4	3057.5
	<i>Nousias17*</i> [10]	2.800	2.287	200.6	56.3
	Proposed	18.311	4.621	176.9	14.6
Optimized	<i>Nousias17</i> [10]	4.063	2.178	621.5	899.4
	<i>Nousias17*</i> [10]	1.165	0.581	245.6	45.9
	Proposed	0.791	0.520	10.8	6.3
Optimized (with Distortion)	Proposed	0.786	0.514	16.3	8.2

TABLE V: RMS blur radius error in pixels for Raytrix dataset for the calibration procedure proposed using the blur radius identified by the detector proposed.

Blur Error [pix]	Overall	Microlens Types		
		Type 1	Type 2	Type 3
Calibration Stage				
Initial	0.424	0.422	0.370	0.480
Optimized	0.404	0.353	0.364	0.494
Optimized (with Distortion)	0.401	0.353	0.359	0.493

reconstruction error decreases by 24.1%. These findings are also observed for the calibration dataset captured using a Raytrix camera and is independent from the set of corners used in the calibration.

Comparing the proposed corner detector with the one proposed by Nousias *et al.* [10], one can see that the proposed corner detector attains smaller reprojection and reconstruction errors in the nonlinear refinement stage. The smallest errors are obtained using the proposed corner detector and the proposed calibration procedure. In this case, the reprojection error attains sub-pixel error in the linear solution and decreases by 43.9% in the nonlinear refinement to 0.53 pixels. On the other hand, the reconstruction error is below 6 mm.

Raytrix Dataset Results. A Raytrix camera is used to obtain images from 10 different poses of a calibration pattern with a 8×6 grid of 48.2×36.2 mm cells. The estimation of the MPC parameters using the calibration procedure proposed is performed using the corners identified with the detector proposed and with the corner detector [10]. The results are compared with the state of the art calibration procedure [10]. As in the synthetic dataset, the results are compared using the reprojection and reconstruction errors for the different stages of the calibration. The results are presented in Table IV.

The corners identified using the proposed detector allow to estimate camera parameters that exhibit consistently smaller errors than the ones obtained by the camera model estimated using the corners identified with the detector [10]. More specifically, the reprojection error reduces by 50.1% and the reconstruction error decreases by 81.3% for the state of the art calibration procedure [10] with the correction defined in Section IV-B. For the calibration proposed, the reprojection error reduces by 34.3% and the reconstruction error decreases by 68.2%. The combination that provides the smallest errors corresponds to the proposed calibration procedure with the

TABLE VI: Parameters of the camera model proposed in Section IV. The parameters are estimated using the calibration procedure proposed and by transforming the parameters (camera model equivalent parameters) estimated using the calibration procedure of Nousias *et al.* [10] according with the mappings defined in Section IV-B. The ratio between the camera model equivalent parameters and the estimated using the proposed calibration is presented in the last row.

Model	k_u	k_v	u_0	v_0	x_0 [m]	y_0 [m]	z_0 [m]	Δu_0	Δv_0	Δx_0 [mm]	Δy_0 [mm]
Proposed	2352.30	2336.40	-257.92	-230.20	0.13	0.11	-1.16	0.90	1.54	0.44	0.77
Bok <i>et al.</i> [8]	2535.38	2535.38	-144.24	-208.09	0.07	0.10	-1.21	0.94	1.64	0.45	0.78
Ratio	1.08	1.09	0.56	0.90	0.54	0.87	1.04	1.05	1.06	1.01	1.02

proposed corner detector, as in the synthetic dataset. In this case, the reprojection error decreases to 0.52 pixels (10.5% decrease) and the reconstruction error decreases to 6.3 mm (86.3% decrease).

The radial distortion present in the Raytrix images acquired is very small, therefore the decrease in the reprojection error with the estimation of radial distortion is only 1.2%. The previous discussions did not include the radial distortion because the state of the art method does not consider distortion estimation during the calibration.

The results in the synthetic and Raytrix dataset show that besides the common extrinsic parameters, the MPC can be described using common intrinsic parameters among the microlens types. More specifically, considering an intrinsic model with 7 parameters (Section III), one is able to obtain smaller or similar reprojection and reconstruction errors with the state of the art calibration procedure [10] that considers 6 parameters for each microlens type in a total of 18 parameters.

Blur Model. In Table V, one presents the blur radius error obtained with the blur model (Section V) used to represent the different microlens types. The blur model parameters are estimated using the calibration procedure and detector proposed. The overall blur error obtained after nonlinear refinement is 0.40 pixels. More specifically, the blur error for the microlens type 1 is 0.35 pixels, for the microlens type 2 is 0.36 pixels and for the microlens type 3 is 0.49 pixels. The sub-pixel blur radius error shows that the blur model is suitable to represent the defocus exhibited by the different microlens types. In Figure 7, one can see that the blur radius estimated is in accordance with the blur radius detected. The blur model gives us a different focal plane depth for each microlens type as expected. Namely, there are two microlenses (types 1 and 2) focusing at depths near the camera (1.19 m and 1.53 m) and one microlens (type 3) focusing at a depth farther away from the camera (2.66 m).

Bok *et al.* [8] Comparison. In Table VI, one presents the parameters obtained for the camera model proposed (Section IV). The parameters are either estimated using the calibration procedure proposed (Section VII) or by transforming the camera model parameters (denoted as camera model equivalent parameters) of Bok *et al.* [8] according with the mappings defined in Section IV-B.

The camera model parameters of Bok *et al.* [8] are obtained using the calibration procedure of Nousias *et al.* [10] and considering the same detected corner points as the ones used in the calibration proposed. In the calibration procedure of Nousias *et al.* [10], one assumes additionally that there is only

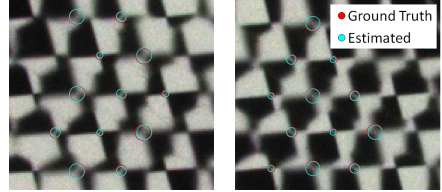


Fig. 7: Examples of the blur radius estimated using the calibration proposed (cyan circles) and comparison with the detected blur radius (red circles).

one microlens type as considered for modeling the MPC. The calibration results in the following additional intrinsic parameters $K_1 = 18.48$ and $K_2 = -22291.00$, and coordinates for converting normalized coordinates to image coordinates are $(f_x, f_y, c_x, c_y) = (46853.00, 46853.00, 2682.10, 3858.00)$.

Additionally, for transforming the Bok *et al.* [8] parameters to the camera model equivalent parameters, one needs to know the origin (p_0, g_0) and the spacing (d_h, d_v) between microlenses. These parameters are obtained by analyzing the microlens centers in the white image during the process of defining the microlens coordinates (k, l) . This analysis gives an horizontal distance of $d_h = 34.89$ pixels and a vertical distance of $d_v = 30.22$ pixels with an origin defined by $(p_0, g_0) = (16.50, 12.63)$ pixels.

The camera model parameters estimated and the camera model equivalent parameters are very similar. Namely, most of the parameters are within a maximum deviation of 10%. The exceptions correspond to the principal point and the (x_0, y_0) coordinates for the origin of the camera coordinate system. The different estimates for these parameters can be caused by the different calibration procedures used and in part can explain the different results in terms of reprojection and reconstruction errors.

IX. CONCLUSIONS

In this work, we proposed a camera model to describe the microlens array of an MPC based on the works of Monteiro *et al.* [14] and Baba *et al.* [15]. This camera model considers an affine mapping to describe the point projections of a world point in each microlens and a blur model to describe the different microlens types' focal planes with a total of 7 intrinsic parameters and $M + 1$ additional blur parameters (M is the number of microlens types).

7. Manuscripts

13

The camera model is used to define a calibration procedure for the MPC based on corner points and blur radius detected in the MIs. The proposed algorithm for feature detection is able to estimate the corner location and the blur radius ensuring robustness to noise thanks to the combination of a local analysis on each MI and a geometry-based refinement across different MIs within each cluster.

The calibration procedure and the corner detector are evaluated on a synthetic and on a dataset acquired with a commercially available MPC. The corner detector algorithm and the calibration proposed outperform the state of the art showing that the MPC can be described using common intrinsic and extrinsic parameters among the different microlens types. In terms of future work, we want to explore how the additional blur information can be used to estimate depth and generate viewpoint-like images for Raytrix.

REFERENCES

- [1] W. Ahmad, L. Palmieri, R. Koch, and M. Sjöström, "Matching light field datasets from plenoptic cameras 1.0 and 2.0," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2018, pp. 1–4.
- [2] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.
- [3] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 43–54.
- [4] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.
- [5] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009, pp. 1–8.
- [6] C. Perwass and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 829 108–829 108.
- [7] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1027–1034.
- [8] Y. Bok, H.-G. Jeon, and I. S. Kweon, "Geometric calibration of microlens-based light field cameras using line features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 287–300, 2017.
- [9] N. Zeller, F. Quint, and U. Stilla, "Calibration and accuracy analysis of a focused plenoptic camera," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 2, no. 3, p. 205, 2014.
- [10] S. Nousias, F. Chadebecq, J. Pichat, P. Keane, S. Ourselin, and C. Bergeles, "Corner-based geometric calibration of multi-focus plenoptic cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 957–965.
- [11] Q. Zhang, C. Zhang, J. Ling, Q. Wang, and J. Yu, "A generic multi-projection-center model and calibration method for light field cameras," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [12] K. H. Strobl and M. Lingenauber, "Stepwise calibration of focused plenoptic cameras," *Computer Vision and Image Understanding*, vol. 145, pp. 140–147, 2016.
- [13] C. Heinze, S. Spyropoulos, S. Hussmann, and C. Perwass, "Automated robust metric calibration algorithm for multifocus plenoptic cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 5, pp. 1197–1205, 2016.
- [14] N. B. Monteiro, J. P. Barreto, and J. A. Gaspar, "Standard plenoptic cameras mapping to camera arrays and calibration based on dlt," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [15] M. Baba, M. Mukunoki, and N. Asada, "A unified camera calibration using geometry and blur of feature points," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 816–819.
- [16] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 972–986, 2012.
- [17] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, 2013.
- [18] O. Johannsen, C. Heinze, B. Goldluecke, and C. Perwaß, "On the calibration of focused plenoptic cameras," in *Time-of-Flight and Depth Imaging*. Springer, 2013, pp. 302–317.
- [19] C. Hahne, A. Aggoun, and V. Velisavljevic, "The refocusing distance of a standard plenoptic photograph," in *2015 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2015, pp. 1–4.
- [20] C. Birkbauer and O. Bimber, "Panorama light-field imaging," in *Computer Graphics Forum*, vol. 33. Wiley Online Library, 2014, pp. 43–52.
- [21] P. David, M. Le Pendu, and C. Guillemot, "White lenselet image guided demosaicing for plenoptic cameras," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*. IEEE, 2017, pp. 1–6.
- [22] S. G. Marto, N. B. Monteiro, J. P. Barreto, and J. A. Gaspar, "Structure from plenoptic imaging," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2017 Joint IEEE International Conference on*. IEEE, 2017, pp. 338–343.
- [23] C. G. Harris, M. Stephens *et al.*, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.
- [24] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 1, pp. 105–119, 2008.
- [25] Y. Bok, H. Ha, and I. S. Kweon, "Automated checkerboard detection and indexing using circular boundaries," *Pattern Recognition Letters*, vol. 71, pp. 66–72, 2016.
- [26] N. B. Monteiro, S. Marto, J. P. Barreto, and J. Gaspar, "Depth range accuracy for plenoptic cameras," *Computer Vision and Image Understanding*, vol. 168, pp. 104–117, 2018.
- [27] N. Zeller, F. Quint, and U. Stilla, "From the calibration of a light-field camera to direct plenoptic odometry," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1004–1019, 2017.
- [28] O. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993.
- [29] C. Liu, S. G. Narasimhan, and A. W. Dubrawski, "Matting and depth recovery of thin structures using a focal stack," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6970–6978.
- [30] Raytrix GmbH, "Raytrix RxLive Software," <https://raytrix.de/downloads/>, accessed on 07.03.2019.
- [31] L. Palmieri, R. Koch, and R. O. H. 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)*. IEEE, 2018, pp. 649–653.
- [32] I. Her, "Geometric transformations on the hexagonal grid," *IEEE Transactions on Image Processing*, vol. 4, no. 9, pp. 1213–1222, 1995.
- [33] H. Lutkepohl, "Handbook of matrices," *Computational Statistics and Data Analysis*, vol. 2, no. 25, p. 243, 1997.
- [34] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, 2000.
- [35] Q.-T. Luong and O. D. Faugeras, "Self-calibration of a moving camera from point correspondences and fundamental matrices," *International Journal of computer vision*, vol. 22, no. 3, pp. 261–289, 1997.
- [36] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [37] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*. Siam, 2000, vol. 1.
- [38] T. Michels, A. Petersen, L. Palmieri, and R. Koch, "Simulation of plenoptic cameras," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2018, pp. 1–4.
- [39] T. Michels, A. Petersen, and R. Koch, "Creating realistic ground truth data for the evaluation of calibration methods for plenoptic and conventional cameras," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 434–442.

Geometric Calibration of Multi-Focus Plenoptic Cameras Supplementary Material

Nuno Barroso Monteiro, Luca Palmieri, Tim Michels, Leandro Cruz,
Reinard Koch, Nuno Gonçalves, José António Gaspar

INTRODUCTION TO THE SUPPLEMENTARY MATERIAL

The supplementary material deduces some of the formulas used in the main paper and provides more insights regarding the camera model and the mapping between the lightfield intrinsic matrix (LFIM) \mathbf{H} and the microlens camera array. Namely, one deduces the camera model of a plenoptic camera (Supp. Material A and B), and one explains the parameterization of the rays and the influence of re-parameterization on the LFIM (Supp. Material C). In Supp. Material D, one presents the location of the microlens projection centers and the restriction to consider the microlens cameras as pinholes. The reduction of the number of non-zero entries in the LFIM by considering the microlens projection centers location is explained in Supp. Material E.

The relationships of the LFIM with the camera models of Zhang *et al.* [1] and Bok *et al.* [2] are presented in Supp. Material F and Supp. Material G, respectively. Additionally, the results provided in Section VIII are segregated by microlens type (Supp. Material H and the focus measure calibration for blur radius detection is further explained in Supp. Material I.

A. Virtual Plenoptic Camera Model

The LFIM was first used to describe a SPC (Figure A.1). More specifically, the camera model proposed by Dansereau *et al.* [3] considers a virtual plenoptic camera whose microlenses define a rectangular tiling. This camera is obtained after a decoding process that transforms the 2D raw image into a 4D LF. This decoding process comprises segmentation of the microlens images (MIs), alignment of the image sensor relatively to the microlens array, and hexagonal sampling correction (Figure A.2.c). For more details, please refer to [3], [4].

[†]This work was supported by the Portuguese Foundation for Science and Technology (FCT) project [grant numbers UID/EEA/50009/2019, and PD/BD/105778/2014] and funded from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

[‡]N. B. Monteiro and J. A. Gaspar are with the Institute for Systems and Robotics, University of Lisbon, Portugal, {nmonteiro,jag}@isr.tecnico.ulisboa.pt

[§]L. Palmieri, T. Michels and R. Koch are with the Department of Computer Science, Kiel University, Germany, {lpa,tmi,rk}@informatik.uni-kiel.de

[¶]L. Cruz and N. Gonçalves are with the University of Coimbra, Institute for Systems and Robotics, and Portuguese Mint and Official Printing Office, INCM Lab, Portugal, lmcruz@isr.uc.pt, nunogon@deec.uc.pt

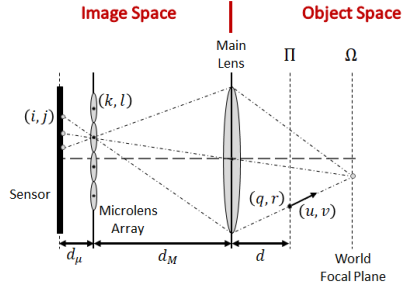


Fig. A.1: Geometry of a SPC considering the microlenses as pinholes and the main lens as a thin lens. The LF in the image space is parameterized using pixels and microlenses indices while the LF in the object space is parameterized using a point and a direction. The LF in the object space can be parameterized on an arbitrary plane Π regardless of the original plane Ω in focus.

The LFIM that describes this virtual camera and maps the rays in the image space to the rays in the object space is obtained by applying a series of six transformations

$$\mathbf{H}_v = \mathbf{H}^{M \rightarrow \Pi} \mathbf{H}^M \mathbf{H}^{S \rightarrow M} \mathbf{H}_m^\phi \mathbf{H}_a^m \mathbf{H}_r^a, \quad (\text{A.1})$$

resulting in a 5×5 homogeneous matrix \mathbf{H}_v with 12 non-zero entries

$$\mathbf{H}_v = \begin{bmatrix} h_{qi} & 0 & h_{qk} & 0 & h_q \\ 0 & h_{rj} & 0 & h_{rl} & h_r \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.2})$$

In this virtual camera, the coordinates (i, k) and (j, l) are independent, and therefore the series of transformations can be analyzed separately for each pair of coordinates without loss of generality. Additionally, note that the pixel coordinates (i, j) are defined relatively to the origin of the corresponding microlens. Thus, starting with the homogeneous coordinates $[i, k, 1]^T$, the transformation

$$\mathbf{H}_r^a = \begin{bmatrix} 1 & N & -n_i \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.3})$$

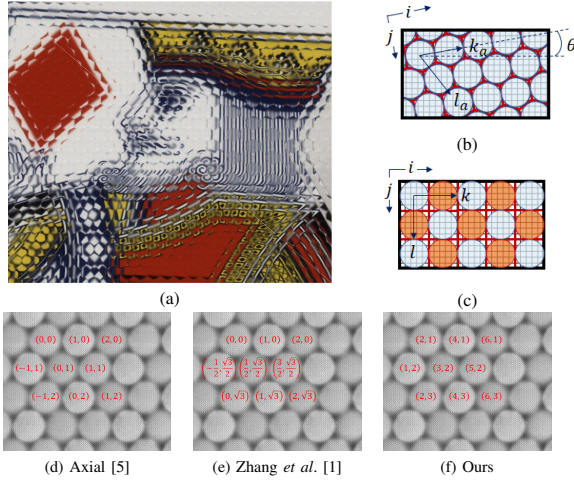


Fig. A.2: Real and virtual microlens array structure of a plenoptic camera. The real microlens array (b) defines an hexagonal tiling that is not aligned with the image sensor by an angle θ and can be represented using an axial coordinate system (d-e) or a cartesian coordinate system (f). The microlens array with the hexagonal structure can be identified in the raw image of an MPC [6] (a). The virtual microlens array (c) created by Dansereau *et al.* [3] defines a rectangular tiling that is aligned with the image sensor. The virtual microlens array is obtained after a decoding process whose rays of the missing microlenses (in orange) are estimated by interpolation.

obtains the 2D image sensor coordinates (pixels in the 2D raw image) associated with a given ray in the image space assuming that each microlens has N pixels and there is a translational pixel offset n_i .

$$\mathbf{H}_a^m = \begin{bmatrix} \frac{1}{f_s} & 0 & -\frac{o_i}{f_s} \\ 0 & \frac{1}{f_k} & -\frac{o_k}{f_k} \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.4})$$

converts the 2D image sensor coordinates and microlenses coordinates (k, l) to metric coordinates by assuming that there are f_c samples per meter and an offset o_c . This allows to define the 4D ray using two points defined in two planes, the image sensor and the microlens array. On the other hand, the mapping

$$\mathbf{H}_m^\phi = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{d_\mu} & \frac{1}{d_\mu} & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.5})$$

allows to change the two-plane parameterization of the ray to a point and a direction defined in the image sensor plane using the distance between the image sensor and the microlens array (d_μ). This parameterization allows to use ray transfer matrices to propagate the ray to an arbitrary plane. Namely,

$$\mathbf{H}^{S \rightarrow M} = \begin{bmatrix} 1 & d_\mu + d_M & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.6})$$

propagates the ray in free space from the image sensor to the main lens and defines the position of the ray in the main lens. d_M is the distance between the microlens array and the main lens. Additionally, the mapping

$$\mathbf{H}^M = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{f_M} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.7})$$

describes the refraction that occurs at the main lens with focal length f_M . This allows to obtain the direction (u, v) in the object space without being modified by the optics of the plenoptic camera. Finally, the transformation

$$\mathbf{H}^{M \rightarrow \Pi} = \begin{bmatrix} 1 & d & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.8})$$

defines the origin of the ray in the object space at a point (q, r) in an arbitrary plane Π at a distance d from the main lens. The six transformations allow to parameterize the ray in metric units by a point in plane Π and a direction.

B. Plenoptic Camera Model

A plenoptic camera has a microlens array with hexagonal tiling that is not aligned with the image sensor (Figure A.2.b). Thus, the camera model for a plenoptic camera should include

the decoding transformations considered for the virtual plenoptic camera. Considering this plenoptic camera, the LFIM that maps the rays in the image space to the rays in the object space is obtained by applying a series of seven transformations

$$\mathbf{H}_r = \mathbf{H}^{M \rightarrow \Pi} \mathbf{H}^M \mathbf{H}^{S \rightarrow M} \mathbf{H}_r^\phi \mathbf{H}_r^m \mathbf{H}_{a\mu}^a \mathbf{H}_r^{a\mu}, \quad (\text{A.9})$$

resulting once again in a 5×5 matrix \mathbf{H}_r , but with 20 non-zero entries instead

$$\mathbf{H}_r = \begin{bmatrix} h_{qi} & h_{qj} & h_{qk} & h_{ql} & h_q \\ h_{ri} & h_{rj} & h_{rk} & h_{rl} & h_r \\ h_{ui} & h_{uj} & h_{uk} & h_{ul} & h_u \\ h_{vi} & h_{vj} & h_{vk} & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.10})$$

The major difference relatively to \mathbf{H}_v (A.1) is on the mapping that allows to obtain the pixel in the raw image associated with a given ray in the image space. This transformation in the virtual plenoptic camera is described using one mapping \mathbf{H}_r^a while in the plenoptic camera this transformation is represented by the product of two mappings, $\mathbf{H}_{a\mu}^a$ and $\mathbf{H}_r^{a\mu}$.

The hexagonal grid of microlenses can be represented by indices using an axial coordinate system (k_a, l_a) [5] whose basis differ from the standard cartesian coordinate system (k, l) (Figure A.2). The transformation between the two different coordinate systems makes the ray coordinates dependent, and therefore the LF coordinates should be analyzed simultaneously. Namely,

$$\mathbf{H}_r^{a\mu} = \underbrace{\begin{bmatrix} 1 & 0 & N_i & 0 & -n_i \\ 0 & 1 & 0 & N_j & -n_j \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}_r^a} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{S}_\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}_s} \quad (\text{A.11})$$

converts the LF coordinates (i, j, k_a, l_a) to pixel coordinates assuming that each microlens has N_i and N_j pixels horizontally and vertically, respectively, and there is a translational pixel offset (n_i, n_j) . The mapping \mathbf{H}_s describes the hexagonal sampling of the microlenses which have a well defined basis (axial coordinate system). Considering the width (N_i) and height (N_j) of the microlenses, the microlens sampling is defined as

$$\mathbf{S}_\mu = \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{bmatrix}. \quad (\text{A.12})$$

Alternatively, one can redefine this basis considering the distance R between the center of the hexagon that includes the MI and the hexagon corners, or considering the horizontal (d_h) and vertical (d_v) distances between consecutive microlenses centers. In these cases, the hexagonal sampling would be defined as

$$\mathbf{S}_\mu = \begin{bmatrix} \sqrt{3} & -\frac{\sqrt{3}}{2} \\ 0 & \frac{3}{2} \end{bmatrix} \quad \text{or} \quad \mathbf{S}_\mu = \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & 1 \end{bmatrix}, \quad (\text{A.13})$$

respectively, where $d_h = N_i = \sqrt{3}R$ and $d_v = \frac{3}{4}N_j = \frac{3}{4}2R$.

Additionally, the misalignment of the microlens array relatively to the image sensor introduces more dependencies among the coordinates of the ray in image space. This misalignment is described by the mapping

$$\mathbf{H}_{a\mu}^a = \begin{bmatrix} \cos \theta & \sin \theta & 0 & 0 & -c_i \\ -\sin \theta & \cos \theta & 0 & 0 & -c_j \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (\text{A.14})$$

that encodes a rotation θ and a translation of (c_i, c_j) pixels of the microlens array relatively to the image sensor. The remaining matrices represent the same transformations described in Supp. Material A. Normally, the misalignment between the microlens array and the image sensor is small, and in recent MPCs can be ignored. Thus, one can consider that $\mathbf{H}_{a\mu}^a = \mathbf{I}_{5 \times 5}$ where $\mathbf{I}_{5 \times 5}$ is a 5×5 identity matrix. This simplifies the mapping \mathbf{H}_r and the resulting matrix has 14 non-zero entries

$$\mathbf{H}_r = \begin{bmatrix} h_{qi} & 0 & h_{qk} & h_{ql} & h_q \\ 0 & h_{rj} & 0 & h_{rl} & h_r \\ h_{ui} & 0 & h_{uk} & h_{ul} & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.15})$$

In the virtual plenoptic camera, the misalignment is also not considered in the camera model. The decoding process that originates the virtual plenoptic camera corrects this misalignment and the microlens hexagonal sampling. Namely, the rectangular tiling of the microlenses [3] result in a sampling that is described by a rectangular basis $\mathbf{S}_\mu = \mathbf{I}_{2 \times 2}$ where $\mathbf{I}_{2 \times 2}$ is the 2×2 identity matrix. This implies that the coordinates of the rays in the image space, (i, k) and (j, l) , are independent (Supp. Material A). However, the transformations associated with the decoding process introduce some aliasing artifacts [4] that may affect calibration and reconstruction results.

In order to maintain the same structure of (A.2) to represent the plenoptic camera, one can incorporate the axial coordinate system basis in the microlenses coordinates as in Zhang *et al.* [1] (Figure A.2.e). Namely, considering the microlens coordinates defined as $[k_z, l_z]^T = \mathbf{S}_\mu [k_a, l_a]^T$ with $\mathbf{S}_\mu = \begin{bmatrix} \sqrt{3} & -\frac{\sqrt{3}}{2} \\ 0 & \frac{3}{2} \end{bmatrix}$. However, this originates non-integer indices which might difficult the access to a particular microlens. Alternatively, one can use a rectangular sampling basis without resorting to a decoding process (Figure A.2.f). This allows to represent the hexagonal structure of the microlens centers in ray image coordinates (p, g) using integer (k, l) coordinates given by $[p, g]^T = \text{diag}(d_h, d_v) \mathbf{S}_\mu [k, l]^T + [p_0, g_0]^T$ where $\mathbf{S}_\mu = \text{diag}(\frac{1}{2}, 1)$ and (p_0, g_0) correspond to the origin for the (k, l) coordinates in the raw image. These approaches allow to model a plenoptic camera with a LFIM \mathbf{H} identical to the one described by Dansereau *et al.* [3] with 12 non-zero entries, *i.e.* $\mathbf{H}_r = \mathbf{H}_v$.

C. Ray Parameterization and Re-Parameterization

In this section, one summarizes the parameterization and re-parameterization of the LF in the object space presented in [7].

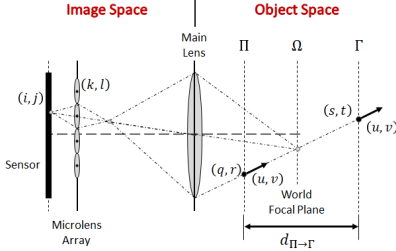


Fig. A.3: Geometry of a FPC. The LF in the image space is parameterized using pixels and microlenses indices while the LF in the object space is parameterized using a point and a direction. The LF in the object space can be parameterized on an arbitrary plane regardless of the original plane Ω in focus.

Let us consider a LF in the object space $L_{\Pi}(q, r, u, v)$ acquired by a plenoptic camera with the plane Ω in focus (Figure A.3). $L_{\Pi}(q, r, u, v)$ is a set of rays, where each ray $\tilde{\Psi}_{\Pi} = [q, r, u, v, 1]^T$ is parameterized using a point (q, r) on a plane Π and a direction (u, v) defined in metric units [8]. The notation (\cdot) represents a vector in its homogeneous coordinates. This LF is mapped to the LF in the image space $L(i, j, k, l)$ by the LFIM \mathbf{H}_{Π} introduced by Dansereau *et al.* [3]:

$$\tilde{\Psi}_{\Pi} = \mathbf{H}_{\Pi} \tilde{\Phi} \quad , \quad (\text{A.16})$$

where $\tilde{\Phi} = [i, j, k, l, 1]^T$ corresponds to a ray that is parameterized by pixels (i, j) and microlenses (k, l) indices and

$$\mathbf{H}_{\Pi} = \begin{bmatrix} h_{qi} & 0 & h_{qk} & 0 & h_q \\ 0 & h_{rj} & 0 & h_{rl} & h_r \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad . \quad (\text{A.17})$$

This mapping allows writing the positions (q, r) and the directions (u, v) as affine mappings on the pixels (i, j) and microlenses (k, l) indices.

On the other hand, the LF in the object space $L_{\Pi}(q, r, u, v)$ can be redefined on another plane Γ by shifting the parameterization plane Π along the optical axis of the FPC, *i.e.* along the normal to the plane Π . Assuming that Γ is at a distance $d_{\Pi \rightarrow \Gamma}$ from Π , the re-parameterization [9] is defined as

$$\tilde{\Psi}_{\Gamma} = \mathbf{D} \tilde{\Psi}_{\Pi} \quad (\text{A.18})$$

where

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & d_{\Pi \rightarrow \Gamma} & 0 & 0 \\ 0 & 1 & 0 & d_{\Pi \rightarrow \Gamma} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad . \quad (\text{A.19})$$

Note that \mathbf{D} maps a ray $\tilde{\Psi}_{\Pi}$ to a ray $\tilde{\Psi}_{\Gamma} = [s, t, u, v, 1]^T$ representing a ray passing through a point (s, t) in plane Γ with a direction (u, v) . Notice that \mathbf{D} changes the camera coordinate system origin but does not change the directions (u, v) .

Mapping the LF in the object space $L_{\Pi}(q, r, u, v)$ to the LF in the image space $L(i, j, k, l)$ by (A.16), one has

$$\tilde{\Psi}_{\Gamma} = \mathbf{D} \mathbf{H}_{\Pi} \tilde{\Phi} \quad . \quad (\text{A.20})$$

The intrinsic matrix $\mathbf{H}_{\Gamma} = \mathbf{D} \mathbf{H}_{\Pi}$ maps the LF in the image space $L(i, j, k, l)$ to the LF in the object space $L_{\Gamma}(s, t, u, v)$.

D. Microlens Pinhole Constraint

In this section, we show that the LFIM [3] can represent an array of parallel microlens cameras. In this work, we follow the same steps of Monteiro *et al.* [7] that demonstrated that the LFIM can represent an array of parallel viewpoint cameras.

From the previous section, let us consider the LF in the object space whose rays are parameterized at a plane Π using a point $[q, r, 0]^T$ and a direction $[u, v, 1]^T$ (Figure A.3). The LFIM \mathbf{H}_{Π} (A.17) maps the rays in the image space $\tilde{\Phi}$ to the rays in the object space $\tilde{\Psi}_{\Pi}$ by (A.16) [3]. For a microlens camera, the microlens coordinates (k, l) are fixed and are considered as parameters. Hence, for a microlens camera, the positions (q, r) and the directions (u, v) are affine mappings only on the pixel coordinates (i, j) , namely

$$\begin{cases} q(i; k, \mathbf{H}_{\Pi}) = h_{qi} i + h_{qk} k + h_q \\ r(j; l, \mathbf{H}_{\Pi}) = h_{rj} j + h_{rl} l + h_r \\ u(i; k, \mathbf{H}_{\Pi}) = h_{ui} i + h_{uk} k + h_u \\ v(j; l, \mathbf{H}_{\Pi}) = h_{vj} j + h_{vl} l + h_v \end{cases} \quad (\text{A.21})$$

where the LFIM \mathbf{H}_{Π} is also considered as a parameter. To simplify the notation, we will not include the parameters (k, l, \mathbf{H}_{Π}) in the following expressions.

A ray captured by a FPC and parameterized by (i, j, k, l) intersects the plane Π at point $\mathbf{p}(i, j) = [q(i), r(j), 0]^T$ with a direction $\mathbf{n}(i, j) = [u(i), v(j), 1]^T$. This allows to define an arbitrary point $\mathbf{c}(i, j, \lambda) = [x, y, z]^T$ along the ray [10] as

$$\mathbf{c}(i, j, \lambda) = \mathbf{p}(i, j) + \lambda \mathbf{n}(i, j) \quad , \quad \lambda \in \mathbb{R} \quad . \quad (\text{A.22})$$

Note that by sweeping the range of (i, j) in (A.22) with $\lambda = 0$, one samples an area of the plane Π through which pass all the microlens imaging rays. In addition, by sweeping (k, l) , one obtains all the microlens cameras, and therefore all rays that can be imaged by the FPC. Finally, sweeping λ , allows representing all world points within the field of view of the FPC.

The location of the projection centers of an optical setup is defined by its caustic surface, which is the loci of singularities

in the flux density [10], [11]. The convergence of the rays captured by a camera at a single point, *i.e.* a unique projection center, is considered a degenerate configuration of the caustic surface (point caustic) [10]. Although there are many techniques to derive the caustic surface, in this work, we will consider the Jacobian method [11].

The caustic surface is defined at the points in the object space where the ray to image mapping (A.22) is singular, *i.e.* the mapping from (i, j, λ) to (x, y, z) is singular. The singularities occur at the set of points where the Jacobian matrix of the transformation does not have full rank, *i.e.* points that make the determinant of the Jacobian vanish $\det(\mathbf{J}(\mathbf{c}(i, j, \lambda))) = 0$. Solving the vanishing constraint one obtains two solutions for λ :

$$\lambda_1 = -\frac{h_{qi}}{h_{ui}} \quad \vee \quad \lambda_2 = -\frac{h_{rj}}{h_{vj}}. \quad (\text{A.23})$$

Replacing λ_1 or λ_2 in (A.22) identifies the caustic profile for the microlens camera. The caustic profile of a single microlens consists of a line with (i) unique (x, z) and variable y components if $\lambda = \lambda_1$ or (ii) unique (y, z) and variable x components if $\lambda = \lambda_2$. In case $\lambda_1 \neq \lambda_2$ the microlens is a non-central camera. The microlens camera corresponds to a central camera, *i.e.* a camera with a unique projection center, if and only if $\lambda_1 = \lambda_2$ which imply the model parameters relation

$$\frac{h_{qi}}{h_{ui}} = \frac{h_{rj}}{h_{vj}}. \quad (\text{A.24})$$

Assuming this constraint and replacing λ in (A.22), expanded by the expressions in (A.21), the location of the projection center for a microlens camera (k, l) is given by

$$\mathbf{p}_c = \begin{pmatrix} h_q - \frac{h_{qi}}{h_{ui}} h_u + k \left(h_{qk} - \frac{h_{qi}}{h_{ui}} h_{uk} \right) \\ h_r - \frac{h_{rj}}{h_{vj}} h_v + l \left(h_{rl} - \frac{h_{rj}}{h_{vj}} h_{vl} \right) \\ -\frac{h_{qi}}{h_{ui}} \end{pmatrix}. \quad (\text{A.25})$$

Furthermore, considering all microlens cameras that can be defined, the LFIM can represent a co-planar grid of equally spaced projection centers. Notice that the microlens coordinates (k, l) only affect the x - and y -components of the projection centers while the z -component of the projections centers is always the same.

E. Reducing the Parameters of the LFIM Parameterization

The LFIM has 12 non-zero entries (A.17) but some parameters can be avoided by choosing an appropriate camera coordinate system origin and considering them on the extrinsic parameters. Namely, Monteiro *et al.* [7] defined a LFIM with 8 non-zero entries by choosing the camera coordinate system origin at the plane containing the viewpoint projection centers. In this section, we will show that a similar representation is obtained by choosing the camera coordinate system origin at the plane containing the microlens projection centers.

Considering the parameterization plane Π (Figure A.3) for the origin of the different rays $\Psi_\Pi = [q, r, u, v, 1]^T$ in the

object space, an arbitrary point is defined as $[x, y, z]^T = [q, r, 0]^T + \lambda [u, v, 1]^T$, $\lambda \in \mathbb{R}$ [10]. The re-parameterization of the rays in the object space to the plane Γ (A.18) corresponds to a shift along the z -axis of the camera coordinate system, which results in $[x, y, z_\Gamma]^T = [s, t, 0]^T + \lambda [u, v, 1]^T$ where $s = q + u d_{\Pi \rightarrow \Gamma}$, $t = r + v d_{\Pi \rightarrow \Gamma}$, and $z_\Gamma = z - d_{\Pi \rightarrow \Gamma}$. Thus, the re-parameterization is redundant with the z -translation of the extrinsic parameters. Assuming that the plane Γ corresponds to the plane containing the microlens projection centers at $d_{\Pi \rightarrow \Gamma} = -h_{qi}/h_{ui}$ (Supp. Material D), one obtains a LFIM \mathbf{H}_Γ with 10 non-zero entries

$$\mathbf{H}_\Gamma = \begin{bmatrix} 0 & 0 & h_{sk} & 0 & h_s \\ 0 & 0 & 0 & h_{tl} & h_t \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.26})$$

Furthermore, extending the definition of the point (s, t) to consider the LF coordinates in the image space and redefining x and y as $x_\Gamma = x - h_s$ and $y_\Gamma = y - h_t$, one obtains $[x_\Gamma, y_\Gamma, z_\Gamma]^T = [h_{sk} k, h_{tl} l, 0]^T + \lambda [u, v, 1]^T$. Hence, the entries h_s and h_t are redundant with the (x, y) -translational components of the extrinsic parameters [7], [3]. Thus, removing the redundant entries, one obtains a LFIM \mathbf{H}_Γ with 8 non-zero entries

$$\mathbf{H}_\Gamma = \begin{bmatrix} 0 & 0 & h_{sk} & 0 & 0 \\ 0 & 0 & 0 & h_{tl} & 0 \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.27})$$

Considering this representation for the LFIM, the microlens projection centers location (A.25) reduces to

$$\mathbf{p}_c = \begin{bmatrix} k & h_{sk} \\ l & h_{tl} \\ 0 & 0 \end{bmatrix}. \quad (\text{A.28})$$

F. Generic 6-Intrinsic-Parameter Model Transformation

The model proposed by Zhang *et al.* [1] considers that the minimal form for the 5×5 LFIM \mathbf{H}_z has 6 non-zero entries

$$\mathbf{H}_z = \begin{bmatrix} 0 & 0 & h_{sk} & 0 & 0 \\ 0 & 0 & 0 & h_{tl} & 0 \\ h_{ui} & 0 & 0 & 0 & h_u \\ 0 & h_{vj} & 0 & 0 & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.29})$$

A similar representation with 6 non-zero entries has been proposed by Marto *et al.* [12] to represent a camera array with co-planar projection centers and whose cameras are identical (same intrinsic parameters). Nonetheless, this normally is not the case for plenoptic cameras, *i.e.* the cameras in the array are not identical [7], [13].

The LFIM defined in Supp. Material E has a minimal form with 8 non-zero entries as a consequence of the different intrinsic parameters between different microlens cameras (Section IV). In this section, one shows that in fact the model

proposed by Zhang *et al.* [1] is equivalent to the 8 non-zero entries representation for the LFIM (A.27) considering that the two additional radial distortion parameters defined in Zhang *et al.* [1] relatively to Brown [14] are included in the \mathbf{H}_z matrix. The two additional parameters for example, in the SPC, are responsible for defining an epipolar plane image (EPI) geometry that is consistent with the zero disparity plane at the main lens world focal plane [7], [8].

Let us consider the relationship between a point $[x, y, z]^T$ in the object space and the distorted ray $\Psi_z = [s, t, u, v]^T$ in the object space as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix} + z \begin{bmatrix} u \\ v \end{bmatrix}. \quad (\text{A.30})$$

Additionally, consider the radial distortion model proposed by Zhang *et al.* [1] assuming that the standard radial distortion correction $\alpha_r = (1 + k_1 r_{uv}^2 + k_2 r_{uv}^4)$ defined by Brown [14] is constant

$$\begin{cases} u^u = \alpha_r u + k_3 s \\ v^u = \alpha_r v + k_4 t \end{cases}, \quad (\text{A.31})$$

where $r_{uv}^2 = u^2 + v^2$, $[k_1, k_2, k_3, k_4]^T$ denotes the distortion vector, and $\Psi^u = [s, t, u^u, v^u]^T$ is the undistorted ray in the object space. Defining $u' = u + s \frac{k_3}{\alpha_r}$ and $v' = v + t \frac{k_4}{\alpha_r}$ to convert the Zhang *et al.* [1] radial distortion model (A.31) to the model defined by Brown [14], and replacing on (A.30), one has $[x, y]^T = [s, t]^T + z[u' - s \frac{k_3}{\alpha_r}, v' - t \frac{k_4}{\alpha_r}]^T$. In order to obtain a relationship of the form $[x, y]^T = [s, t]^T + z[u', v']^T$, let us define the mapping between the rays in the object space $\Psi = [s, t, u', v']^T$ and Ψ_z as

$$\tilde{\Psi} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{k_3}{\alpha_r} & 0 & 1 & 0 \\ 0 & \frac{k_4}{\alpha_r} & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}_{\Psi_z}^{\Psi}} \tilde{\Psi}_z. \quad (\text{A.32})$$

Extending the definition of Ψ_z to consider the rays in the image space $\Phi = [i, j, k, l]^T$ using the LFIM \mathbf{H}_z proposed by Zhang *et al.* [1], one obtains a LFIM $\mathbf{H}'_z = \mathbf{H}_{\Psi_z}^{\Psi} \mathbf{H}_z$ defined with 8 non-zero entries

$$\mathbf{H}'_z = \begin{bmatrix} 0 & 0 & h_{sk} & 0 & 0 \\ 0 & 0 & 0 & h_{tl} & 0 \\ h_{ui} & 0 & h_{sk} \frac{k_3}{\alpha_r} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} \frac{k_4}{\alpha_r} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (\text{A.33})$$

G. Microlens Camera Array Mapping to LFIM

In this section, one shows that the model proposed by Bok *et al.* [2] can be represented by a LFIM by constraining the microlens centers coordinates on the raw image to be regularly spaced. The structure of the LFIM will depend on

the sampling basis that is considered for representing the microlens coordinates (Supp. Material B).

The microlens camera model, defined by Bok *et al.* [2] and adapted by Nousias *et al.* [15] to describe an MPC, represent the projection of a point in the object space on a microlens using 6 parameters and the knowledge of the microlens center coordinates on the raw image by

$$\begin{bmatrix} \Delta p \\ \Delta g \end{bmatrix} = \frac{1}{K_1 z + K_2} \begin{bmatrix} f_x x - z \hat{p}_c \\ f_y y - z \hat{g}_c \end{bmatrix}, \quad (\text{A.34})$$

where K_1 and K_2 are additional intrinsic parameters to the conventional pinhole camera model [16], $\Delta p = p - p_c$ and $\Delta g = g - g_c$ with (p_c, g_c) defining the microlens center coordinates associated with the raw image coordinates (p, g) , and $\hat{p}_c = p_c - c_x$ and $\hat{g}_c = g_c - c_y$. The (f_x, f_y, c_x, c_y) are the parameters used to convert normalized coordinates to image coordinates. This model can be rewritten to get a pinhole-like representation by isolating the coordinates of the point $[x, y, z]^T$. This allows to define the projection matrix \mathbf{P}_b for the microlens camera (p_c, g_c) as

$$\mathbf{P}_b = \underbrace{\begin{bmatrix} \frac{f_x}{K_1} & 0 & -\frac{p_c - c_x}{K_1} \\ 0 & \frac{f_y}{K_1} & -\frac{g_c - c_y}{K_1} \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K}_b} \underbrace{\begin{bmatrix} \frac{K_2}{K_1} \frac{p_c - c_x}{f_x} \\ \frac{K_2}{K_1} \frac{g_c - c_y}{f_y} \\ \frac{K_2}{K_1} \end{bmatrix}}_{\mathbf{E}_b} \quad (\text{A.35})$$

where \mathbf{K}_b and \mathbf{E}_b correspond to the intrinsic and extrinsic matrix for the microlens camera, respectively. The extrinsic matrix \mathbf{E}_b allows to define the position of the microlens camera array relatively to the camera coordinate system origin which corresponds to the plane containing the viewpoint projection centers [2]. For considering the relationship with the world coordinate system, one should consider the matrix defined as $\mathbf{P}_b {}^c \mathbf{T}_w$ where ${}^c \mathbf{T}_w = \begin{bmatrix} {}^c \mathbf{R}_w & {}^c \mathbf{t}_w \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix}$ defines the rigid body transformation between the world and camera coordinate systems with rotation ${}^c \mathbf{R}_w \in SO(3)$ and translation ${}^c \mathbf{t}_w \in \mathbb{R}^3$, and $\mathbf{0}_{1 \times 3}$ is the 1×3 null matrix. Representing the raw image coordinates by the 4D coordinates of the rays in the image space using $i = \Delta p$, $j = \Delta g$, and (k, l) using the rectangular sampling proposed $\mathbf{S} = \text{diag}(\frac{d_h}{2}, d_v)$ (Figure A.2.f)

$$\begin{bmatrix} p_c \\ g_c \end{bmatrix} = \mathbf{S} \begin{bmatrix} k \\ l \end{bmatrix} + \begin{bmatrix} p_0 \\ g_0 \end{bmatrix}, \quad (\text{A.36})$$

the intrinsic and extrinsic matrices can be redefined as

$$\mathbf{K}_b = \begin{bmatrix} \frac{f_x}{K_1} & 0 & -\frac{p_0 - c_x}{K_1} - k \frac{1}{2} \frac{d_h}{K_1} \\ 0 & \frac{f_y}{K_1} & -\frac{g_0 - c_y}{K_1} - l \frac{d_v}{K_1} \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{A.37})$$

and

$$\mathbf{E}_b = \begin{bmatrix} \frac{K_2}{K_1} \left(\frac{p_0 - c_x}{f_x} + k \frac{1}{2} \frac{d_h}{f_x} \right) \\ \frac{K_2}{K_1} \left(\frac{g_0 - c_y}{f_y} + l \frac{d_v}{f_y} \right) \\ \frac{K_2}{K_1} \end{bmatrix}, \quad (\text{A.38})$$

respectively. The parameters d_h and d_v correspond to the horizontal and vertical distance between consecutive microlenses centers.

The LFIM associated with the camera model of Bok *et al.* [2], considering the plane containing the microlenses projection centers as the origin of the camera coordinate system defined as

$$\mathbf{H}_{\Pi}^b = \begin{bmatrix} 0 & 0 & \frac{K_2}{K_1} \frac{1}{2} \frac{d_h}{f_x} & 0 & \frac{K_2}{K_1} \frac{p_0 - c_x}{f_x} \\ 0 & 0 & 0 & \frac{K_2}{K_1} \frac{d_v}{f_y} & \frac{K_2}{K_1} \frac{q_0 - c_y}{f_y} \\ \frac{K_1}{f_x} & 0 & \frac{1}{2} \frac{d_h}{f_x} & 0 & \frac{p_0 - c_x}{f_x} \\ 0 & \frac{K_1}{f_y} & 0 & \frac{d_v}{f_y} & \frac{q_0 - c_y}{f_y} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (j)$$

The LFIM \mathbf{H}_{Π}^b is not represented in its minimal form the entries non-dependent on the ray coordinates in the image space are not included in ${}^c\mathbf{T}_w$, *i.e.* there are some redundant parameters with the extrinsic parameters. Since the camera coordinate system origin of Bok *et al.* [2] is defined on the plane containing the viewpoint projection centers, one should re-parameterize the rays assuming $d_{\Pi \rightarrow \Gamma} = -K_2/K_1$ (Supp. Material C). This allows to define a LFIM with a minimum of 8 non-zero entries

$$\mathbf{H}_{\Gamma}^b = \begin{bmatrix} -\frac{K_2}{f_x} & 0 & 0 & 0 & 0 \\ 0 & -\frac{K_2}{f_y} & 0 & 0 & 0 \\ \frac{K_1}{f_x} & 0 & \frac{1}{2} \frac{d_h}{f_x} & 0 & \frac{p_0 - c_x}{f_x} \\ 0 & \frac{K_1}{f_y} & 0 & \frac{d_v}{f_y} & \frac{q_0 - c_y}{f_y} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad (\text{A.40})$$

where $[-K_2/f_x, -K_2/f_y, 0]^T$ defines the baseline shift between consecutive viewpoint cameras.

H. Corner Detection: Additional Results

A further analysis on the error of the corner detection. Since the MPCs use three different lens type, we can divide the measurements for each lens type.

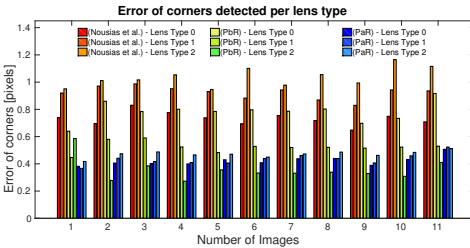


Fig. A.4: Error of the methods

The measurements show some interesting results. While for the approach used in [15] the third lens type seems to exhibit lower error, the same seems to be more difficult to estimate for our method, before the refinement step. In fact, that is the only

measurement for which our approach does not perform better. After the refinement, however, this trend is re-established and the error related to the third lens type is again the lowest.

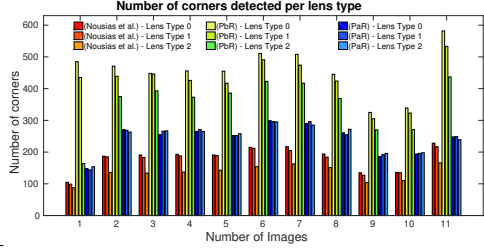


Fig. A.5: Number of corners found from each methods

The first two lens types do not show particularly interesting results and are quite similar to each other.

With respect to the number of corners found, we performed a similar analysis. Here the only visible trend seems to be a lower number of corner found for the first lens type, while the second and the third show similar behaviours.

The meaning behind these results may be related to the amount of focus blur and to particular characteristics of the optics of each lens. Such investigation goes out of the scope of this work and is left for future research.

I. Corner Detection: Blur Calibration

Here a more detailed overview over the blur calibration process is given. The challenge discussed here is intrinsically related to the definition of the blur radius. Even for a human observer is hard to define the blur radius of a blurred corner, so we describe our approach to overcome it.

To ensure consistency and have a clear definition of blur, we make use of synthetic images. After creating an ideal image of a corner, we use the Gaussian Blur plug-in from GIMP to blur the corner of a desired amount. In our experiment, we have 21 different images with a blur radius ranging from 0 to 10 pixels each 0.5 pixels. To emulate real image we also add gaussian noise using Matlab's *imnoise* function, with zero mean and a small variance. In our experiment the variance value was set to be $\sigma_{noise} = 0.0003$.

The focus measure operator chosen for this work was, as explained in the main text, the Tenengrad Variance [17], and it is used to compute a mapping function that, given the focus measure as input, returns the blur radius in pixels.

This procedure is done selecting a window around the corner of a specific size, the same size that it will be later used in the actual blur estimation step. In our experiments we have a window size of 9. The window size is a dataset dependent parameters: the minimum requirement is that the actual values is larger than the maximum blur visible in a MI, to correctly compute it. The largest the window gets, however, will negatively apply to the detected corners that lie at the border of the MIs, since there is not enough information there to accurately calculate the focus measure on that region.

7. Manuscripts

8

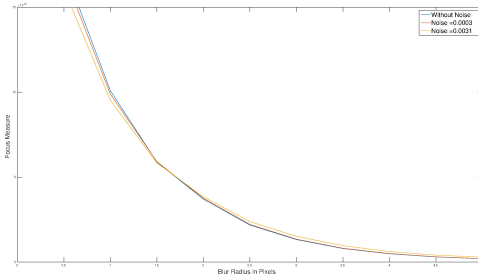


Fig. A.6: The mapping function. On the y-axis, the focus measure is displayed, while on the x-axis the correspondent blur radius in pixel is shown.

Different type of corner template accounting for lens border regions can be taken into account for future research, yet were not considered as a requirement because the parameter has shown to work correctly on our datasets.

REFERENCES

- [1] Q. Zhang, C. Zhang, J. Ling, Q. Wang, and J. Yu, "A generic multi-projection-center model and calibration method for light field cameras," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [2] Y. Bok, H.-G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 287–300, 2017.
- [3] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenslet-based plenoptic cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1027–1034.
- [4] P. David, M. Le Pendu, and C. Guillemot, "White lenslet image guided demosaicing for plenoptic cameras," in *Multimedia Signal Processing (MMSp), 2017 IEEE 19th International Workshop on*. IEEE, 2017, pp. 1–6.
- [5] I. Her, "Geometric transformations on the hexagonal grid," *IEEE Transactions on Image Processing*, vol. 4, no. 9, pp. 1213–1222, 1995.
- [6] W. Ahmad, L. Palmieri, R. Koch, and M. Sjöström, "Matching light field datasets from plenoptic cameras 1.0 and 2.0," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-COIN)*. IEEE, 2018, pp. 1–4.
- [7] N. B. Monteiro, J. P. Barreto, and J. A. Gaspar, "Standard plenoptic cameras mapping to camera arrays and calibration based on dlt," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [8] N. B. Monteiro, S. Marto, J. P. Barreto, and J. A. Gaspar, "Depth range accuracy for plenoptic cameras," *Computer Vision and Image Understanding*, vol. 168, pp. 104–117, 2018.
- [9] C. Birklbauer and O. Bimber, "Panorama light-field imaging," *Computer Graphics Forum*, vol. 33, no. 2, pp. 43–52, 2014.
- [10] M. D. Grossberg and S. K. Nayar, "The raxel imaging model and ray-based calibration," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 119–137, 2005.
- [11] D. G. Burkhard and D. L. Shealy, "Flux density for ray propagation in geometrical optics," *JOSA*, vol. 63, no. 3, pp. 299–304, 1973.
- [12] S. G. Marto, N. B. Monteiro, J. P. Barreto, and J. A. Gaspar, "Structure from plenoptic imaging," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2017 Joint IEEE International Conference on*. IEEE, 2017, pp. 338–343.
- [13] N. B. Monteiro and J. A. Gaspar, "Generalized camera array model for standard plenoptic cameras," in *Iberian Robotics conference*. Springer, 2019, pp. 3–14.
- [14] D. C. Brown, "Decentering distortion of lenses," *Photogrammetric Engineering and Remote Sensing*, 1966.
- [15] S. Nousias, F. Chadebecq, J. Pichat, P. Keane, S. Ourselin, and C. Bergesles, "Corner-based geometric calibration of multi-focus plenoptic cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 957–965.
- [16] O. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993.
- [17] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, 2013.

7.2 Manuscript 2

Analyzing Disparity and Compression across Plenoptic Cameras

Luca Palmieri, Waqas Ahmad, Mårten Sjöström and Reinhard Koch

Submitted to

*Signal Processing Image Communication Special Issue on Light Field Imaging,
2021*

The following manuscript is the preprint version submitted to the Signal Processing Image Communication Special Issue on Light Field Imaging. Personal use of this material is permitted. Additional permissions must be obtained for all other uses, in any current or future media, including reprinting/re-publishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Analyzing Disparity and Compression across Plenoptic Cameras

Luca Palmieri, Waqas Ahmad, Mårten Sjöström and Reinhard Koch

Abstract—Light Field (LF) constitutes a solid foundation for immersive visual applications due to its ability to capture and represent angular and spatial information of the scene. Plenoptic cameras uses microlens array to record the scene on a two-dimensional sensor. The optical setting of plenoptic cameras regulates different multiplexing of LF information, introducing new challenges in the processing. Fewer assumptions can be made on the structure of the input light field, thus novel approaches have to be developed. In addition, capturing LF contents requires huge computational power, large network and storage resources. This work presents a detailed analysis on practical aspects of plenoptic image processing like capturing, rendering, compressing and estimating the disparity on plenoptic images acquired with available commercial implementations. Two of the most important processing tasks, disparity estimation and compression, are investigated across different plenoptic camera models. Using a publicly available dataset, standard and focused plenoptic cameras are compared. Results show that image compression is more beneficial for images captured using focused plenoptic cameras. When compressing the same scene in different representations, the sub-aperture image is more efficient compared to its lenslet representation for both plenoptic camera models. Diverse disparity estimation methods based on different representations are analyzed to reveal their strength and limitations for each camera model. Moreover, the behaviour of disparity under different compression rates is inspected, giving further insights about possible applications.

Keywords—compression, disparity, light field, plenoptic camera, representation

NOTATION

In this work, the following notation will be used:

- sub-aperture image (SAI) refers to a conventional viewpoint image rendered from a captured light field;
- microlens image (MI) refers to an image captured within a single microlens on the sensor of a plenoptic camera; the image formed on the sensor by all MIs is denoted as lenslet image;
- standard plenoptic camera (SPC) refers to plenoptic cameras where the microlens array (MLA) is placed at the focal plane of the main lens, also denoted as conventional plenoptic cameras, described in [1];
- focused plenoptic camera (FPC) refers to plenoptic cameras where the MLA is moved away from the focal plane of the main lens, described in [2].
- spatial and angular resolution refers to their pixel properties: spatial resolution refers to the width and height of the rendered image for the angular resolution to the number of perspective views for the sub-aperture images;
- angular density refers to the distance in pixels between two consecutive angular sample.

I. INTRODUCTION

Light Field (LF) imaging has recently received significant research attention mainly due to the large availability of LF contents. The field is growing into several branches, ranging from acquisition,

compression, to visualization of the captured contents. LF acquisition technologies offer various possibilities to sample the spatial and angular information of the scene. Initially, the LF information was captured using a set of calibrated and synchronized multiple traditional cameras [3]. Later on, thanks to the advancements in optical technology, handheld plenoptic cameras were introduced [1], [4]. This enabled capturing the LF in a single shot, reducing the hardware costs and computational complexity, while allowing the capture of dynamic scenes.

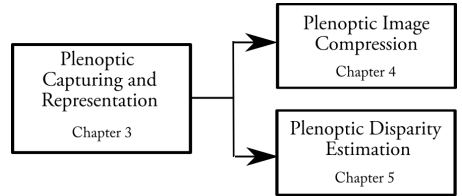


Fig. 1: A scheme to summarize the contents and the structure of the paper for an easier visualization.

Plenoptic camera records the spatial and angular information of the scene onto a single two-dimensional image by placing a MLA between the main lens and image sensor. The arrangement of main-lens, MLA and image sensor regulates the sampling of the LF. Plenoptic cameras has demonstrated their influence in various applications, e.g. refocusing, scene reconstruction, depth estimation and novel view synthesis. Since its introduction, the resolution of the images captured with plenoptic cameras has seen a dramatic improvement, at the cost of additional transmission and storage resources. The traditional image and video encoders can be used to compress plenoptic image with limited compression efficiency. The recent initiative of JPEG referred to as JPEG Pleno [5] to seek efficient compression and grand challenges [6], [7] for plenoptic image compression highlights the importance of LF compression.

Parallel to the hardware improvement, software algorithms are also proposed to supplement the rendering of captured LF information. Since the LF exhibits redundancy, its cues are used to calculate a super resolved rendered image. Several approaches have been proposed in this direction, trying to maximize the optical capability using focused plenoptic cameras and dedicated rendering models to generate images with higher spatial resolution [8], computing simultaneous depth and all-in-focus super-resolved images [9], analyzing and restoring the images using bayesian framework [10], [11], or using a variational approach and makes use of disparity maps to render novel views [12], [13]. Although different in their implementations, the methods have a common background: the knowledge of geometrical information about the scene, mostly in form of a disparity or depth map. Therefore, the analysis of different disparity estimation methods assumes particular importance. Disparity methods showed promising results using stereo matching on microlens images [14], [15], building a cost volume using the sub aperture images [16]–[21] and analyzing the images in the epipolar plane image domain [22], [23].

Contributions

The scope of this work is to give an overview of the different plenoptic camera configurations and to investigate two of the most critical processes involved in LF imaging, namely compression and disparity estimation. More in details, the main contributions of this paper are:

- 1) an overview of the characteristics of the different plenoptic camera models and their respective commercial implementation in Section III.
- 2) an improved version of the rendering algorithm for focused plenoptic cameras capable of deliver high resolution images incorporating several steps in the image formation process in Section III-D¹,
- 3) a novel analysis on the compression methods, where we introduced the comparison using different encoders for the same scene captured with both standard and focused plenoptic cameras in Section IV.
- 4) a comparison of disparity estimation approaches based on different lightfield (LF) representations acquired with both plenoptic camera models in Section V,
- 5) an analysis of the disparity estimation under different compression rates in Section V-E.

Structure of the paper

The rest of the paper is structured as depicted in Figure 1: in Section II the state of the art is discussed, in Section III the capturing and representation process with different plenoptic cameras is discussed, in Section IV the task of light field compression is analyzed, in Section V the challenges of disparity estimation are covered and finally in Section VI conclusions are presented.

II. STATE OF THE ART

The first standard plenoptic cameras were introduced in [1] and made commercially available by the Lytro company [25]. They achieved capturing of a four dimensional light field in a single shot, yet resolution limitations were evident. The standard plenoptic camera saw a second generation with higher resolution [25] allowing LF processing to be treated as a special case of multi-view stereo, with most approaches exploiting the SAIs representation [18], [21]. Further work compared and evaluated images acquired with standard plenoptic cameras with images from a synthetic dataset generated with a virtual camera at different position, equivalent to a camera array with a narrow base line around 18 – 20 micro meters [26].

Capturing and Representations

The development of focused plenoptic camera had the larger improvement in terms of spatial resolution [8], [27]. The trade-off in their work consists, assuming a fixed microlens size, in reducing the angular density to increase the spatial resolution. Because of the low resolution of the original standard plenoptic camera and the advances in novel view synthesis, this trade-off is beneficial for many application [27]. The introduction of the focused plenoptic cameras accounted for some modification in the rendering process. Contrary to the initial theoretical explanation [27], as noticed in [28], the rendering of all-in-focus images when capturing the scene with a focused camera requires depth information of the scene, as investigated in [9], [29]. This requirement poses specific constraints to the image model formation, particularly the challenge of processing the microlens images *before* the rendering. Such a shift is visible in several work related to filtering [30], calibration [31] or depth estimation [32]. The focused plenoptic camera instead became more of a niche topic, and its commercial implementation from the Raytrix

company [33] introduced an additional novelty, namely the three different focal lengths in the microlens array [4]. Such cameras go under the name of multi-focus plenoptic cameras. Having three different focal lengths enables multi-focus plenoptic cameras to enlarge the depth of field of the camera by a factor of three, while reducing the spatial resolution only by the half [4]. However, the computational effort increased significantly to account for microlens images with different amounts of blur. Recently, novel approaches have been proposed including calibration [31], toolbox for disparity estimation [15] and additional softwares to render realistic synthetic images [24].

Compression

Due to availability of benchmark contents [34] and grand challenges [6], [7], compression of plenoptic images captured using a SPC as the Lytro camera received significant attention from the research community compared to plenoptic images captured using FPC like the Raytrix camera. Initially, the lenslet representation of plenoptic images was exploited by using modified HEVC intra coding schemes [35]–[37]. Later the compression efficiency was improved by using video coding schemes on the SAI representation [38], [39].

Disparity estimation

As previously mentioned, estimating the disparity is one of the main applications of LF processing. Therefore, many approaches have been proposed to deal with this challenge. Some methods work directly on the lenslet image [14], [15] building a cost volume based on the pixel similarity and using stereo-matching to exploit the redundancy of the MI to calculate disparity. Another approach is to combine correspondence and focal cues [16] or specifically modelling more sophisticated characteristics as the shadows [17] or the occlusions [19]. An iterative refinement function was also used to reach higher accuracy in the estimation in [18]. The estimation has been addressed also in the epipolar plane image (EPI) domain with variational framework [22] and building a dedicated operator to detect slopes of the lines [23]. Recently, learning-based approaches started to show promising results, learning pixel-based similarity measures [20] or training a neural network on the stack of SAI to extract features and achieve accurate estimations [21].

III. PLENOPTIC CAPTURING AND REPRESENTATION

In this section we review the different possibilities for capturing a scene with MLA-based plenoptic cameras and the various representation and the relative trade-off in terms of resolutions. The terms MLA-based plenoptic camera refers to a camera where a MLA is placed between the sensor and the main lens. Each microlens samples a portion of the scene, allowing the recording of the light field in a single shot. In fact, based on their position and their size, the microlenses collects a different set of light rays from the scene. The optical configuration of the main lens, MLA and sensor regulates the sampling process of the scene.

A. Camera Models

The first camera model we want to describe is the SPC. This camera model, described graphically in Figure 2a, is able to capture the light field with high density of angular information. Its optical configuration places the MLA at the focal plane of the main lens. Because of this choice, each microlens capture only one spatial point and its angular information, as visible in Figure 2d. The advantage of this approach is the possibility of rendering a large number of viewpoints image from a single shot. On the other hand, its limitation consists in the spatial resolution, that amounts exactly to the number of microlenses in the camera [27].

To overcome the spatial resolution limitation, the second camera model was introduced, namely the FPC, depicted in Figure 2b. The difference consists in the placement of the MLA, that is now shifted

¹The code relative to the rendering algorithm has been released under the GPL license as part of the Plenoptic Toolbox, available at the following link: <https://github.com/ftreafki/PlenopticToolbox2.0>

7. Manuscripts

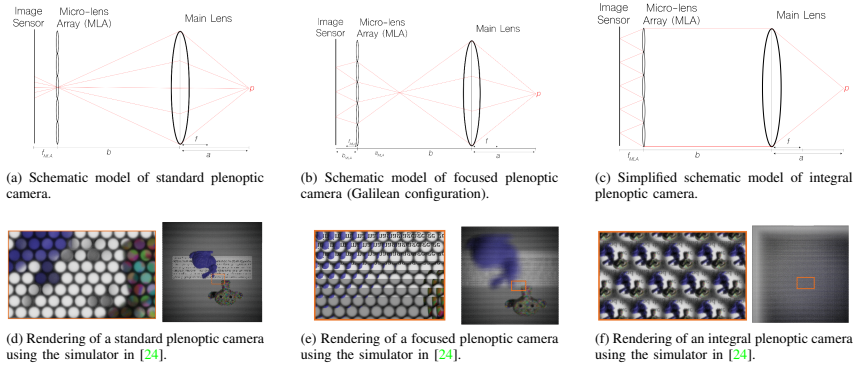


Fig. 2: Schematic model and image rendering for the three camera models explained in the text. Focused plenoptic cameras have two possible configurations, Galilean and Keplarian. Here the Galilean was chosen for an easier visualization. The two configurations are equivalent. The rendering are obtained using the plenoptic simulator described in [24].

so that its focal plane coincide with the one of the main lens. This means that in each MI a portion of the scene is captured, similarly to a camera-array, as visible in Figure 2e. It follows that the spatial resolution of the rendered image increases, since each single MI records not only the spatial information of one point, but several of them. The price to pay in this case is the reduction of the angular density in the captured light field.

If we were to follow this intuition and set up the camera so that each MI records the whole scene, the recorded spatial resolution would increase to reach exactly the MI size, while strongly reducing the angular density. In fact, if each lens capture the whole scene, the angular resolution corresponds to the number of microlenses, similarly to the spatial resolution in the case of SPC. We denote such a camera model integral plenoptic camera (IPC) and it is the third model that we take into account. Although originally developed before the other camera models, it was discontinued and never reached a commercial implementation, therefore it is less widely used. There are different possible implementations as result of combination of main lens, prisms, negative or positive lenses, as well described in [2]. Another approach has been taken in [40] and [41], where a more sophisticated system of lenses is used to recreate similar condition for light field microscopy. A simplified schematic model of this camera model is graphically depicted in Figure 2c and one virtually generated image is shown in Figure 2f.

The images shown in Figures 2d, 2e and 2f were not acquired with plenoptic cameras, instead were generated using the plenoptic simulator described in [24]. They are intended to give a visual explanation about how each camera model capture the same scene.

B. Commercial Implementations

In this section we aim at analysing the differences between the different plenoptic camera models. To compare several applications on images captured with actual plenoptic cameras, we rely on the commercial implementations. While the simulator can be fine-tuned to achieve almost all possible combinations of light field sampling, commercial implementations impose more rigid constraints.

Even though the first mention of light field dates to almost a century ago, MLA-based plenoptic cameras flourished only recently and did not yet became a standard camera for the consumer market, therefore commercial implementations are limited. The most used

plenoptic cameras are the Lytro [25], now discontinued, built following the SPC model. The Raytrix camera [33] instead constitutes a special case of the FPC model, namely the multi-focus plenoptic cameras. Both cameras are used for research purposes, and Raytrix cameras extended their usage to industrial application. The prototype of the third camera model was produced by Adobe, as described in [2], but was never made commercially available. It can be seen in Figure 3. Since there are no commercially available cameras of the third type for IPC, this model received less attention and therefore will not be used for this work.



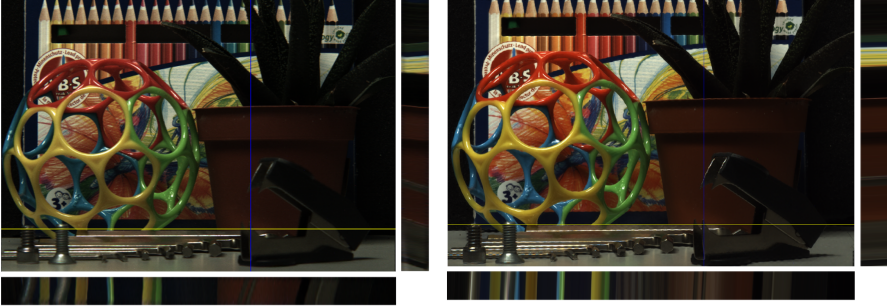
Fig. 3: Commercial implementations of plenoptic cameras.

To ensure a fair comparison between plenoptic camera models we have to rely on images acquired from cameras produced by Lytro and Raytrix. In this work we use images from the database acquired using Lytro Illum and Raytrix R29 [42], the plenoptic cameras with highest quality available. Lytro's Illum capture raw images of 5368×7728 pixels, while Raytrix R29 raw images consists of 4384×6576 pixels.

C. On the Angular-Spatial Resolution Trade-off

Each camera models samples the light field with a different pattern. As already pointed out in previous works [2], an increase in the spatial resolution results in a decrease in the angular resolution, and vice versa. The three models account for the whole range of LF sampling, from lowest to highest angular and spatial resolution, and are suitable to give a full overview of the trade-offs of LF capturing.

If we assume a fix sensor size and a similar structure with regards to the MLA size, we can summarize the characteristics of the different camera in a very simple and intuitive way in Table I. We denote with MI the microlens size and with $N_{L,h}$ and $N_{L,v}$ respectively



(a) Scene acquired with a Lytro Illum Camera and rendered with the toolbox in [43]

(b) Scene acquired with a Raytrix R29 Camera and rendered with the proposed algorithm.

Fig. 4: The central sub-aperture image and two slices in the epipolar plane from a scene acquired with both Lytro and Raytrix cameras. [42]

the number of microlenses in the horizontal and vertical direction. Spatial and angular resolution are denoted respectively as ϕ and ρ . Since their value is camera dependent, the interval $(1, MI)$ is defined.

The three camera models describe the whole spectrum of the spatio-angular trade-off. The SPC and the IPC can be considered as the two bounds for the resolutions. When spatial resolution drops to the minimum, angular resolution reach its maximum. No gain in terms of overall resolution is achieved with a different optical configuration, only the sampling distribution is varied. For the case of SPC and IPC, it is straightforward to calculate the actual size of the rendered image if we assume no post-processing. It is dependent on the number of lenses and the MI size, or in other terms, the number of pixels in the sensor behind a single microlens.

Camera Model	Resolution for MI		Resolution for rendered image	
	Spatial (ρ)	Angular (ϕ)	Spatial (ρ)	Angular (ϕ)
SPC	1	MI	$N_{L,h} \times N_{L,v}$	MI
FPC	$\rho, \phi \in (1, MI)$		depth-dependent	
IPC	MI	1	MI	$N_{L,h} \times N_{L,v}$

TABLE I: Summary of the spatio-angular resolution trade off for different plenoptic camera models. ρ refers to spatial and ϕ to angular resolution. Values in pixels.

The FPC is a more complex case, where the actual resolution depends on the depth of a single point in the scene. As described in [28], [44], the spatial resolution increases when the image projected through the main lens is closer to the MLA, and the angular density follows the opposite path.

Because of the characteristics of the different plenoptic cameras, a careful setup is required to obtain a similar light field representation when the same scene is captured with two different cameras. To achieve this we propose an improved version of the rendering algorithm, which enables a larger flexibility in the parameterization and allows to create a more unified representation, a requirement for a fair comparison between the cameras.

D. Rendering Process

Our aim is to create a unified representation to compare images taken with different plenoptic cameras, namely SPC and multi-focus plenoptic camera (MPC). Using two different camera models results in a different sampling of the light field. However, the recorded information is partially redundant and can be represented in the same

way. The chosen solution is to adopt the SAI representation, which constitutes the standard for most application and serves as basis for the EPI domain. An example is visible in Figure 4.

Rendering SAIs from a LF captured with SPC can be done by selecting one pixel from each MI, and this method is implement in the LF Toolbox [43] which was used to render the images. Rendering SAIs from a LF captured with FPC involves the extraction of a patch instead of one single pixel for each MIs. This process requires the disparity information, because, as discussed in Section III-C, the spatial resolution is not constant across the image. The patch size is directly proportional to the spatial resolution and depends on the position of each object in the three dimensional scene, i.e. its distance from the camera. So in the following, we assume the disparity has been estimated. For more information about the disparity estimation, please refer to Section V.

Two approaches have been proposed to estimate the patch sizes and tile them together to generate a SAI [28], [29]. The first proposed algorithm in [28] assumes all microlenses have the same focal length, which is not the case for the multi-focus plenoptic cameras. In general, both approaches [28], [29] use integer pixels for the window size, reducing the quality of the rendered image.

We propose an adapted version of the rendering algorithm to overcome these limitations: it handles different types of microlenses without rendering blurred part of the image and work on patch sizes with floating precision numbers. The underlying strategy remains the same: for each lens a patch size is selected based on the disparity information available, and those patches tiled together generates the final image. When dealing with different microlens types, we need a technique to merge seamlessly the MI with different blur levels. This is achieved by creating one image for each microlens type, dividing regions based on the disparity values. When merging the images for the rendering, instead of discarding the information recorded on the unfocused microlenses, we use a weighted average, where the weights for the focused part are significantly larger, avoiding any shifts or artefacts.

The first step of our algorithm consists in the classification of the microlenses. This is done by exploiting the regular grid structure of the MLA that, as analyzed in [14], [45], follows the same pattern. The information about the microlenses focal length is known only to manufacturers, yet using the RxLive software [46] to capture the images, a configuration file with information about the microlens ranges can be extracted. Alternatively, a calibration can be performed to estimate the optical configuration of the microlenses, or a classification based

7. Manuscripts

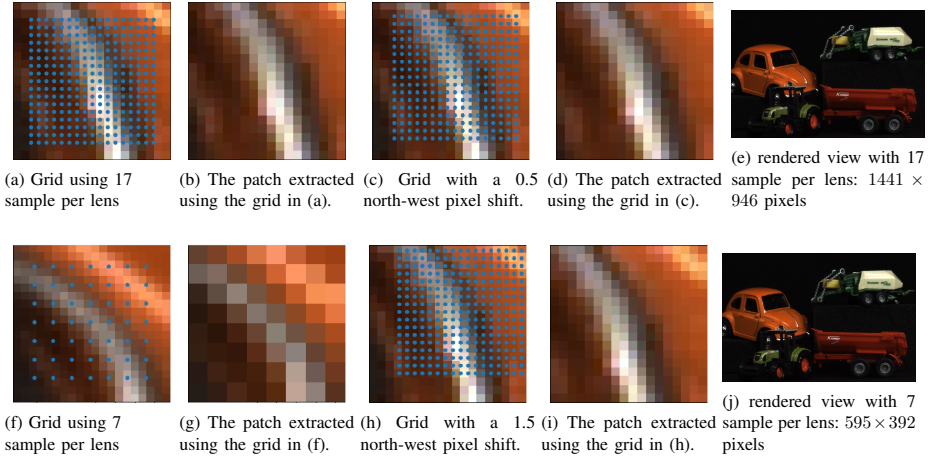


Fig. 5: The images shows excerpts of the rendering algorithm. Please notice how up-scaling and down-scaling are integrated due to the choice of the sample per lens. Also the shift can assume decimal values, allowing us to create more view with less disparities. This is particularly important when creating views as input for different algorithms. The two views have very different resolution, yet the shows no visual difference, proving the robustness of our algorithm. No visible artifacts arises at larger resolutions.

on focus measure can also be applied, as in [31].

The second step consists in the extraction of the patches from each MI. The patch size is calculated based on the disparity value and the camera configuration. Instead of using a window of $n \times n$ ($n \in \mathbb{Z}$) pixels, a regular sampling grid of size $ps \times ps$ ($ps \in \mathbb{R}$) is applied. For a more accurate extraction, the raw image is interpolated, which enables the usage of decimal values to select the optimal patch size and the distance between each sample within the grid. The distance between samples is calculated from the number of samples per lens. By tuning this parameter, one can control the final spatial resolution of the rendered image, as shown in Figure 5. This sampling mechanism implicitly performs the down- and up-sampling usually required for the rendering. Because the number of samples per lens is constant across different microlenses, larger patch size will have larger distance between samples, thus resulting always in a patch of the desired size, without need of resizing.

The proposed algorithm allows a more flexible generation of different viewpoint images. Since the approach work on interpolated image, the constraints on the integer pixel do not exist anymore and we can move the viewpoint by a desired amount. By doing this, we can control the parallax between adjacent views and we can generate denser or sparser sets of SAIs. This allows us to recreate a number of SAIs similar to the SAIs obtained from an image captured with a Lytro camera by changing the parameters in the rendering algorithm.

The rendering algorithm also has its limitations. The maximum parallax in the scene cannot be increased, only the density of the sampling can be chosen. Moreover, it requires the microlens centers grid to be known and the estimation of the disparity to be correct. However, the rendering algorithm is robust to noise due to the way the estimation is handled. First, for every microlens only one patch size is needed, more specifically only one disparity value. This enable the usage of approaches that calculates a coarse disparity map, as described in [14], [47], which reduces the noise in the estimation.

Second, disparity estimation, particularly stereo-matching, tends to fail in texture-less regions while being more accurate in areas full of features. This results in accurate patch size for highly textured region, where the exact patch size is needed to avoid artifacts, and less accurate estimation of the patch size in uniform regions, where the patch size is almost irrelevant for the rendering. Therefore the rendering image does not show any artifact. This allows to render artifacts-free viewpoint images with sparse or noisy disparity images, which would be significantly harder in the case of a geometrical approach that projects points in the three dimensional space and then trace them back to render the image.

An example of the same scene captured with Lytro and Raytrix plenoptic cameras is visible in Figure 4. The two rendered images looks very similar, although the plenoptic image captured with the Raytrix camera has higher spatial resolution and lower angular resolution. In Figure 5 some example of the rendering algorithm and two version of the same scene rendered with low and high spatial resolution are shown to support our explanation.

IV. PLENOPTIC IMAGE COMPRESSION

The plenoptic image captured using Lytro camera contains angular correlation within each microlens and spatial correlation among neighbouring microlenses. In Raytrix R29 camera, each microlens captures a small portion of the scene. The angular correlation exist among neighbouring microlens and spatial correlation is contained within each microlens. As discussed in the previous section, a plenoptic image can be transformed into its SAI representation, which shows the different perspectives of the scene. Firstly, a compression analysis is presented for Lytro and Raytrix plenoptic images in lenslet format using image coding standard. Secondly, a comparison is presented between lenslet representation and SAI representation using image and video coding standard.

A. Selected Dataset

The research is conducted using publicly available dataset [42]. The dataset contains real scenes captured using the Lytro Illum camera and the Raytrix R29 camera from same view position and under the same conditions. The table II shows the selected LF images from the dataset.

TABLE II: Selected LF images from dataset in [42].

S/N	Image Name	Resolution (WxH)
1	RTX 007	6576x4384
2	RTX 015	6576x4384
3	IMG 001	7728x5368
4	IMG 004	7728x5368

B. Performance Comparison

For the performance comparison, the peak signal to noise ratio (PSNR) is used as a quality measure and calculated using the original and the compressed image, as described in Equation 1:

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \quad (1)$$

where the mean square error (MSE) is calculated by:

$$MSE = \frac{1}{WH} \sum_{x=1}^W \sum_{y=1}^H [I(x, y) - I'(x, y)]^2 \quad (2)$$

The rate is calculated in bits per pixel (bpp) and is obtained by dividing the compressed size by the total number of pixels of a raw plenoptic image.

C. Plenoptic image compression using image encoder

The High Efficiency Video Coding (HEVC) coding standard has shown significant compression efficiency compared to its predecessor Advanced Video Coding (AVC) and JPEG2000 [48]. The HEVC image encoder mainly relies on its intra prediction modes to exploit the spatial correlation present in the data. HEVC uses DC, Planar and 33 directional modes for prediction of the current block from already encoded neighboring pixels [48]. Fig. 6 shows an example of HEVC intra prediction modes, i.e. vertical, horizontal, diagonal and DC.

In recent past, researchers has proposed modifications in HEVC image coding standard to improve the rate distortion (RD) efficiency for plenoptic images [35], [36]. However, such changes requires modification in reference HEVC standard which already deployed on most of the available software and hardware infrastructures. An experiment is performed to evaluate the coding efficiency for Lytro and Raytrix plenoptic images using already available HEVC reference coding scheme. The plenoptic images captured using Lytro and Raytrix cameras are given as input to reference HEVC image coding scheme and compressed on four different bit-rates in order to cover different bit rate scenarios. The compression efficiency is evaluated based on RD relationship. Fig. 7 shows the RD performance of HEVC-intra coding for Raytrix and Lytro plenoptic images.

The Raytrix plenoptic images show significantly better compression efficiency compared to Lytro plenoptic images. The HEVC-Intra coding traverses the plenoptic image block by block and takes prediction for each current block from already encoded neighbouring blocks. The neighboring border pixels are simply replicated on the current block. Figure 8 shows the small portion of Lytro and Raytrix plenoptic images. In Lytro plenoptic image, pixels under each microlens have different intensity values which is not very suitable for HEVC Intra prediction modes. However, in Raytrix plenoptic image each microlens records small portion of scene that may contain uniform regions. Hence, the HEVC-intra prediction modes works better on Raytrix plenoptic images compared to Lytro plenoptic image.

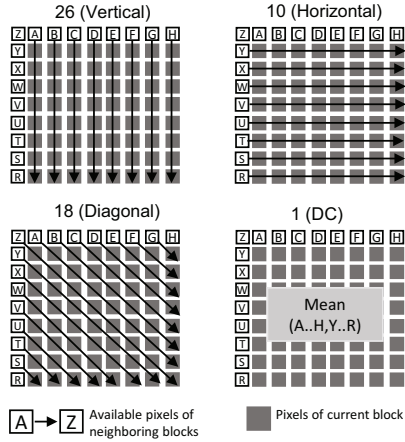


Fig. 6: Example of Intra prediction available in HEVC. The pixels of current block are interpolated using border pixels mark with labels A-H and R-Z.

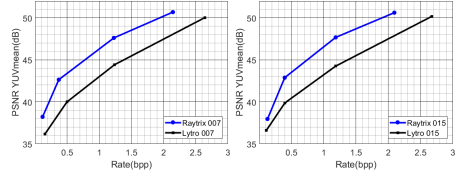


Fig. 7: Rate Distortion Analysis between Raytrix and Lytro plenoptic images

D. Plenoptic image compression using video encoder

Each SAI depicts the scene from a unique perspective which mimics the views of multiple camera system captured with narrow baseline. High correlation exists among neighbouring SAIs that can be efficiently coded using MV-HEVC based coding scheme [49]. The MV-HEVC scheme converts the two dimensional set of SAIs into multiple pseudo video sequences. The coding scheme enables each SAI to take two dimensional prediction from its neighbouring SAIs. Fig. 9 shows the coding structure of MV-HEVC scheme for 5×5 SAIs which is interpreted as five pseudo videos, each of them with five frames.

1) *Lytro*: The 13 multi-view sequences with each having 13 frames are given as input to MV-HEVC encoding scheme. The Fig. 10 shows the rate distortion performance between MV-HEVC based coding scheme applied on SAIs in comparison with HEVC intra based lenslet coding. It can be seen that MV-HEVC based coding scheme perform significantly better than HEVC intra based coding. The SAI representation of plenoptic image makes the input more suitable for video encoder which is built with assumptions of image and video signal. The state-of-the-art tools in video coding efficiently encodes the SAIs.

2) *Raytrix*: The optical configuration of Lytro camera enables a very simple methodology for transforming plenoptic image into SAIs

7. Manuscripts

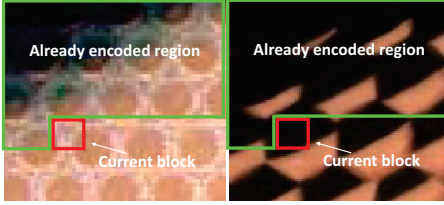


Fig. 8: A visualization of the compression of a block for images captured respectively using a Raytrix and a Lytro camera when using the HEVC-Intra coding scheme.

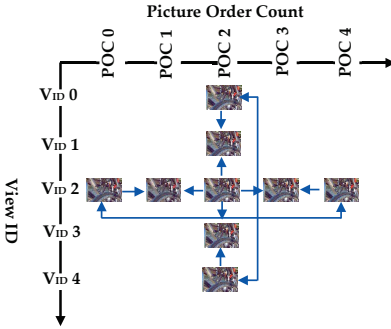


Fig. 9: MV-HEVC based coding scheme for a set of 5x5 sub-aperture images.

since by gathering a single pixel from each micro lens generates a single perspective of the view. On the other hand, the optical setting of Raytrix camera complicates the SAI generation process. The depth information is required to accurately generate the SAIs from plenoptic image. In this paper we have generated SAIs from Raytrix plenoptic image using four different sampling methods, i.e., using either all microlenses or using only focused microlenses. The table III shows the angular and spatial resolution of each SAI generation scheme.

TABLE III: Sub-aperture image representation for plenoptic images captured with a Raytrix R29 Camera.

S/N	Scheme	Angular Resolution	Spatial Resolution
1	All microlens	13x13	1612x1054
2	Focused microlens	13x13	1062x690
3	Focused microlens	7x7	1062x690
4	All microlens	5x5	1612x1054

Figure 11 shows the compression efficiency of each SAI generation scheme when given as input to MV-HEVC coding scheme [49]. In all the SAI generation schemes the total number of pixels of raw plenoptic image (6576×4384) are used as reference for estimating the bits per pixel for each scheme. For both LF images it can be observed that the SAI representation with angular resolution of 7x7 provide better compression efficiency. The SAI scheme with 7x7 angular resolution produces less redundant information in SAIs and the sampling of pixels from focused lenses results in better quality

138

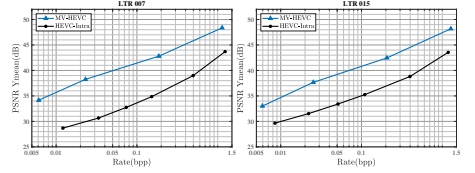


Fig. 10: Compression of Lytro lenslet images using the HEVC Intra coding scheme and sub-aperture images using the MV-HEVC coding scheme.

of SAIs.

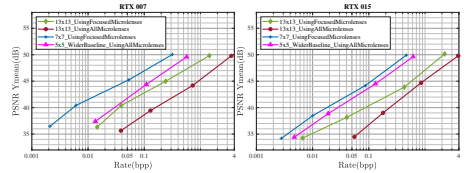


Fig. 11: Sub-aperture image captured with a Raytrix camera compressed using the MV-HEVC scheme.

The Fig. 12 shows the comparison between lenslet coding and SAI coding of Raytrix plenoptic image. It can be seen that SAI representation of Raytrix plenoptic image yields better compression efficiency compared with lenslet coding.

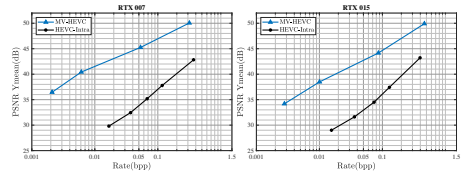


Fig. 12: Compression of images captured with a Raytrix camera using the MV-HEVC coding scheme for sub-aperture images and the HEVC intra coding scheme for the lenslet images.

V. PLENOPTIC DISPARITY ESTIMATION

At this point, it is clear that the disparity estimation plays a crucial role in the processing plenoptic image. As expected, there are several possibilities that have been investigated to find the optimal solution for this task. Each approach has its benefits and limitations. In this chapter we give an overview about the different methods for estimating the disparity of an image acquired with a plenoptic camera. We divide the approaches into three main categories based on the representation chosen: microlens images (MI), sub-aperture images (SAI) and epipolar plane images (EPI). The analysis is performed on images selected from the dataset in [42], which allows us to investigate the same scene captured with Lytro and Raytrix plenoptic cameras.

























Image	Disparity with SPC			Disparity with MPC			
	SAI [18]	SAI [21]	EPI [23]	MI [15]	SAI [18]	SAI [21]	EPI [23]
							
							
							

Fig. 13: Overview of disparity and depth estimation methods using different representation and approaches. Each method was developed for a specific input type of images, yet all methods can recover the overall shape of the scene. Nevertheless, some methods clearly shows much noisier results. Best viewed in colors.

A. Using Microlens Images (MIs)

The first category includes approaches that estimate disparity from the information recorded in each MI. Datasets suitable for these approaches have been made available in [5], [42]. In the FPC and MPC case, this can be seen as a special case of multi-view stereo, and correspondence matching can be applied. Images acquired with SPC show unfocused MI and do not need geometrical information to transform the sampled light field to the SAI domain, therefore no methods for estimation based on the MI have been proposed, up to our knowledge.

Even though the environment is similar to the standard stereo matching, the size of the MIs introduces new challenges: disparities are low and no information about larger structure can be inferred from a single MI. On the other hand, the advantage in the search for correspondences in the MIs lies in their redundancy. The number of time the same object is imaged in different MI is depth-dependent, yet multiple correspondences are guaranteed, enabling a more robust matching process.

The approaches shown in the literature tend to converge to a similar solution, tackling the delicate process of micro-lenses selection [45], creating a cost volume and extracting the disparity values after a regularization step [14], [15]. Another possibility is the calculation of a feature-based sparse disparity map and a successive filling step to obtain a dense disparity image [50].

In this work we use the method from [15], which delivers a dense disparity map per microlens, to compare disparity estimation based on MIs. The algorithm selects the optimal combination of adjacent microlenses to calculate the cost volume using a pixel-based similarity function. The cost volume is filtered using semi-global matching, disparity labels are extracted and a Taylor expansion is applied to create a continuous version of the disparity map. We denote this disparity as *per lens disparity* to distinguish it from the *conventional* disparity map. We denote as *conventional* disparity the disparity image of a SAI, which allows for a easier comparison with other approaches that use a different representation. In fact, the rendering algorithm described in Section III-D is used to convert the per lens disparity maps to conventional disparity maps for the comparison, applying only a gentle median filter as a post-processing step.

B. Using Subaperture Images (SAIs)

Here we discuss methods based on estimating disparity from SAIs. For images captured with a SPC, the transformation to the SAIs representation is straightforward and does not require knowledge about the geometry of the scene. Moreover, the SAI representation depicts multiple perspective of the captured scene, which is similar to images acquired with camera arrays and therefore more suitable for many

standard applications. A wider access to datasets [26], [42], [51]–[54] and benchmarks [55], [56] accounted for its diffusion, making this approach the most widely used within light field community.

The disparity estimation based on SAI is in fact similar to the idea just discussed for the MI case, being both a special case of multi-view stereo. The difference lies in the structure of the images and the cost volume. A grid of sub aperture images has the advantage of having the same number of views in both direction, and the possibility to use off-the-shelf methods for the correspondences matching, including methods based on a global optimisation on the image structure. Different focal cues, as defocus and correspondences, can be combined to create a more robust cost volume, as presented in [16] and extended in [17] where the shading is also taken into account. The usage of global optimization techniques lead to more accurate result, as evident with the solution adopted in [18] that makes use of graph cuts and iterative methods to refine the cost volume. Other methods focused on handling occlusion [19] or proposed a learning-based approach, as learning the most effective pixel similarity measure [20] or training a network to calculate the disparity based on a subset of the SAI [21].

Creating SAIs from the same scene captured with a SPC and a MPC and using them as input for one method allows to obtain an interesting view on how the same approach perform on different sampling of the light field. In this work two implementations are adapted and used. The one from [18] to estimate disparity maps from SAIs based stereo cues and global optimisation, and the approach described in [21] which uses neural network trained on synthetic images from the light field benchmark [26]. In both cases, parameters have been tuned for optimal results and the set of SAI were created with a similar configuration.

C. Using the Epipolar Plane Image (EPI)

The third representation, also widely used in the light field community, is the EPI. The four-dimensional structure relies on the two-plane parameterization described in [57]. Extracting two-dimensional slices from the structure allows us to analyse the structure of the scene in different dimensions. An example of such slices is shown in Figure 4. The slices are extracted by fixing one axis in the image plane, while collecting the pixels from the angular domain. Therefore, each point in the image becomes a line in the EPI representation. The slope of the line is due to the shift in the rendering from different viewpoints and thus related to its three-dimensional position in the scene.

Using this representation, the aim of the estimation is to calculate the slope of each lines to recover the depth information. A robust estimation requires high angular density and resolution to obtain an image where lines and regions can be detected. A low angular density in a captured plenoptic image causes lines to be separate into segments, and low resolution results in blocky lines, in both cases

preventing an accurate detection and estimation of the disparity of the scene. It follows that scenes acquired with a SPC are more suitable for this technique, because of their larger angular resolution. However, we try to replicate a similar angular resolution in the scenes captured with FPC by selecting a lower parallax between adjacent views.

Different approaches have been proposed to guarantee robustness of the estimation: a local analysis using structure tensor is followed by a global optimization step in [22], a dedicated spinning parallelogram operator is devised in [23].

In this work the spinning parallelogram operator described in [23] was chosen for the comparison. The parameters were adjusted to fit the scene and achieve the best outcome. The operator separates the EPI in two regions based on the lines slope and is able to robustly handle occlusions.

D. Performance Analysis and Comparison

The main challenge in the performance analysis is the lack of ground truth. It is not trivial to gather geometrical information with high accuracy from the same perspective of the scene which was captured with multiple cameras. Synthetic scenes have been used as an alternative solution in recent approaches [26], yet they do not take into account the diversity of the input and the different ways of sampling the light field, they assume always an ideal grid of virtual cameras capturing the scene.

In this work we perform a visual analysis of the results obtained from applying different approach to the same scene. The comparison between the same scene acquired with different camera models gives us a further insight about advantages and limitations of plenoptic camera models. Moreover it allows us to compare methods with very different approach starting from the same recorded light field.

In Figure 13 we show an overview of the results obtained on three images of the dataset [42]. Since we perform a visual analysis, images with particular characteristics were chosen to ensure variability in the estimation. In particular, we are interested in the challenging regions for disparity estimation: specularities (first image), light reflections on metal components (last image), complex structures with thin components (first and last image) and low-texture regions (second image).

Looking at the comparison, it is clear that there is not a single algorithm that outperforms the other in all cases, instead each method shows different behaviours with their advantages and limitations.

The disparity computed on the MI reaches a high accuracy overall, yet it exhibits noisy behaviours and do not deal well with specularities, as noticeable in the first image. This can be explained from the lack of a global refinement step, since the disparity map is calculated for each MI and tiled together, with only a gentle median filter as post-processing. Adding a refinement step on the rendered image may improve the disparity in terms of smoothness and noise reduction. Nevertheless, this approach recovers very well the fine details and the small structures. This is clearly visible in all images, where smaller details are preserved.

The disparity computed on the SAJ using the techniques described in [18] has the refinement step based on the structures in the scene. The optimization uses graph cuts which reduces the noise in the disparity at the price of decreasing the accuracy of fine structures. Even though parameters were tuned to find the optimal trade-off between preserving high frequencies and noise reduction, fine details are lost. This is visible on disparity maps calculated using images captured with the SPC. Surprisingly, the images captured using MPC constitute a better input for the algorithm, and the results preserve the scene structure more reliably.

Using the neural network from [21] brings to slightly different results. Here again no strong post-processing filters were used and therefore the scene is noisier since the network was trained on synthetic images and therefore expected images without noise. Nevertheless, fine structures are very well estimated and using images acquired with the SPC results in very accurate and detailed disparity maps.

Results using the spinning parallelogram operator from [23] show very accurate disparity maps with fine structures preserved and precisely estimated. It seems to work better for the images acquired with SPC and delivers the best results on the regions with specularities, as the dark red ball in the first image.

Overall it can be concluded that each algorithm has its strength and limitations, yet some show more promising results. For the images acquired with SPC, the neural network approach from [21] and the spinning parallelogram operator from [23] both show impressive results, with the latter exhibiting particular strength on the preservation and estimation of finer structures and the former excelling in estimating uniform or textureless regions with very low noise. In the case of FPC however, estimating disparity based on the MI similarity as described in [15] seems to be the most effective approach to reduce the noise and estimate scene details.

E. Evaluation of disparity estimation on microlens images under different compression rates

In this section we investigate the performance of the disparity estimation on microlens images compressed with the HEVC scheme under different compression rates. Since the analysis is performed on MI, our choice is to use images acquired with a FPC. Therefore, for this particular section images acquired with a Raytrix camera were used. Images comes from the database acquired in [42], as for the previous sections.

The conditions for this analysis are the following: working on lenslet images, a disparity estimation method which calculates the disparity for each pixel using a similarity based function is applied [15], therefore we expect a degradation of the quality of the disparity map under lower compression rates. Even though a single algorithm is used, it is safe to generalize the results across different approaches based on matching which relies on pixel similarity. For completely different approaches, the same conclusions may not held. To quantitatively evaluate the image quality, the PSNR measure has been applied on both color and disparity images in this analysis.

The objective of this analysis is to find out whether additional information about different behaviours of the disparity estimation under different compression rates can be beneficial for the choice of the color for different applications.

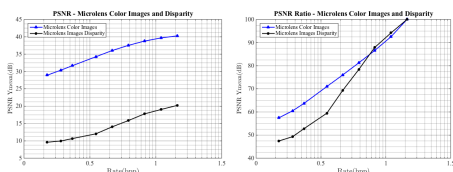


Fig. 14: Analysis of the quality of lenslet image and disparity under different compression rates. On the left absolute values in PNSR are plotted, while on the right the ratio between the PSNR value of a compressed image and the uncompressed image is shown.

The results confirm our hypothesis: a reduction in the quality of the lenslet image due to lower compression rates translates into a degradation of the quality of the estimated disparity maps. This is straightforward and clearly visible in the left graph of Figure 14.

Although both curves look very similar, they have slightly different shapes. This is more visible in the right graph of Figure 14. The curve is a plot of exactly the same data using a different representation. The y-value in this case is the reference ratio: the ratio between the PSNR value of a compressed image and the uncompressed image. Therefore the value is normalized and expressed as a percentage between 0 and

100%. This normalization allows us to see the difference in the shape of the two curves.

At high compression rates, the disparity estimation seems to be less affected from the degradation of the quality of the lenslet image. This effect is relatively small, yet is interesting, because it gives additional reasons for the choice of the optimal compression rates. An explanation could be found in the noise reduction effect of the compression. Slight compression acts as a pre-processing filter and improves the matching in the similarity-based estimation, thus it does not reduce the accuracy of the estimation. However, when the bitrates drops below a certain threshold, in our experiments around 0.6 bpp, the quality of the disparity starts to decrease drastically. This effect is associated with the failure in the correspondence matching caused by a stronger degradation of the lenslet image. This is expected and accounts for a disadvantage of using lower compression rates.

VI. CONCLUSION

The analysis reported in this work aims at expanding the research field to include different types of camera models to account for a wider application range. In recent years, plenoptic images captured using Lytro camera has received enormous research attention compared to plenoptic image captured with Raytrix camera. Various factors contributed to the popularity of Lytro cameras in the research community, e.g., availability of benchmark datasets, availability of Matlab toolbox and its usage in various plenoptic image compression competitions. Initial proposals tend to code plenoptic image in lenslet format by introducing novel tools in HEVC intra coding. Later on, its sub-aperture image representation was exploited using video coding tools which has shown significance compression improvement. The results show that the sub-aperture image representation of Raytrix plenoptic image also perform better using video coding tools compared to lenslet based coding.

Using a similar representation allowed us to set up a fair comparison of disparity estimation techniques from images acquired with both plenoptic camera models. It is possible to obtain high quality results with both plenoptic camera models, yet each camera model has its advantages and limitations. Recent approaches involving a learning step or a dedicated parallelogram operator achieved the most accurate results when using standard plenoptic cameras. Using focused plenoptic camera instead, the best results were achieved by working directly on the microlens image with a matching approach. Moreover, analyzing the behaviour of the disparity estimation under different compression rates delivers further insights on the advantages and limitations of potential use of these images for certain types of applications. This analysis may serve as a basis for future research on the topic.

ACKNOWLEDGMENT

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

REFERENCES

- R. Ng, M. Levoy, B. Mathieu, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.
- G. Todor, Z. K. Colín, C. Brian, S. David, N. Shree, and I. Chintan, "Spatio-angular resolution tradeoffs in integral photography," *Rendering Techniques*, vol. 2006, pp. 263–272, 2006.
- M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.
- P. Christian and W. Lennart, "Single lens 3d-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII*, vol. 8291. International Society for Optics and Photonics, 2012, p. 829108.
- L. Palmieri, "The plenoptic toolbox 2.0," <https://github.com/PlenopticToolbox/PlenopticToolbox2.0>, [Online, 2018].
- M. Kerabek, T. Bruylants, T. Ebrahimi, F. Pereira, and P. Schelkens, "Icme 2016 grand challenge: Light-field image compression," *Call for proposals and evaluation procedure*, 2016.
- C. for Proposals on Light Field Coding, "Jpeg pleno," *ISO/IEC JTC 1/SC29/WG17/4014, 74th Meeting, Geneva, Switzerland*, January 15–20, 2017.
- T. Georgiev and A. Lumsdaine, "Superresolution with plenoptic camera 2.0," *Adobe Systems Incorporated, Tech. Rep.*, 2009.
- F. P. Nava and J. Luke, "Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera," in *2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2009, pp. 1–4.
- T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009, pp. 1–9.
- T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 972–986, 2012.
- S. Wanner and B. Goldluecke, "Spatial and angular variational super-resolution of 4d light fields," in *Computer Vision—ECCV 2012*. Springer, 2012, pp. 608–621.
- , "Variational light field analysis for disparity estimation and super-resolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 606–619, 2014.
- O. Fleischmann and R. Koch, "Lens-based depth estimation for multi-focus plenoptic cameras," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 410–420.
- P. Luca, K. Reinhard, and V. R. O. Het, "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)*. IEEE, 2018, pp. 649–653.
- M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 673–680.
- M. W. Tao, R. Ramamoorthi, J. Malik, and A. A. Efros, "Unified multi-cue depth estimation from light-field images: Correspondence, defocus, shading, and specularities," Ph.D. dissertation, University of California, Berkeley, 2015.
- H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1547–1555.
- T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3487–3495.
- H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon, "Depth from a light field image with learning-based matching costs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- C. Shin, H.-G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4748–4757.
- S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 41–48.
- S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Computer Vision and Image Understanding*, vol. 145, pp. 148–159, 2016.
- T. Michels, A. Petersen, L. Palmieri, and R. Koch, "Simulation of plenoptic cameras," in *2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, June 2018, pp. 1–4.
- Lytro, "Home," <https://www.lytro.com>, [Online, 2018].
- K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision (ACCV)*, 2016.
- A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009, pp. 1–8.
- T. Georgiev and A. Lumsdaine, "Reducing plenoptic camera artifacts," *Comput. Graph. Forum*, vol. 29, pp. 1955–1968, 2010.
- S. Wanner, J. Fehr, and B. Jähne, "Generating epi representations of 4d light fields with a single lens focused plenoptic camera," in *International Symposium on Visual Computing*. Springer, 2011, pp. 90–101.

7. Manuscripts

LIGHT FIELD IMAGING, 2021

11

- [30] T. E. Bishop and P. Favaro, "Plenoptic depth estimation from multiple aliased views," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 1622–1629.
- [31] S. Nousias, F. Chadebecq, J. Pichat, P. Keane, S. Ourselin, and C. Bergeles, "Corner-based geometric calibration of multi-focus plenoptic cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 957–965.
- [32] O. Fleischmann and R. Koch, *Lens-Based Depth Estimation for Multi-focus Plenoptic Cameras*. Springer International Publishing, 2014, pp. 410–420. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11752-2_33
- [33] Raytrix, "Light field technology," <https://www.raytrix.de>, [Online, 2018].
- [34] V. Vaish and A. Adams, "The new stanford light field archive, [online]," <http://lightfield.stanford.edu/lfs.html>, accessed on 2018-08-01.
- [35] C. Conti, P. Nunes, and L. Soares, "Hevc-based light field image coding with bi-predicted self-similarity compensation," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.
- [36] R. Monteiro, L. Lucas, C. Conti, P. Nunes, N. Rodrigues, S. Faria, C. Pagliari, E. da Silva, and L. Soares, "Light field hevc-based image coding using locally linear embedding and self-similarity compensated prediction," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.
- [37] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.
- [38] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.
- [39] W. Ahmad, R. Olsson, and M. Sjöström, "Interpreting plenoptic images as multi-view sequences for improved compression," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4557–4561.
- [40] M. Martínez-Corral, J. Barreiro, A. Llavador, E. Sánchez-Ortega, J. Sola-Pikabea, G. Scrofanì, and G. Saavedra, "Integral imaging with fourier-plane recording," in *Three-Dimensional Imaging, Visualization, and Display 2017*, vol. 10219. International Society for Optics and Photonics, 2017.
- [41] G. Scrofanì, J. Sola-Pikabea, A. Llavador, E. Sanchez-Ortega, J. Barreiro, G. Saavedra, J. Garcia-Sucerquia, and M. Martínez-Corral, "Fimic: design for ultimate 3d-integral microscopy of in-vivo biological samples," *Biomedical optics express*, vol. 9, no. 1, pp. 335–346, 2018.
- [42] W. Ahmad, L. Palmieri, R. Koch, and M. Sjöström, "Matching light field datasets from plenoptic cameras 1.0 and 2.0," in *2018 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-COIN), Stockholm-Helsinki-Stockholm, 3-5 June 2018*, 2018.
- [43] D. G. Dansereau, O. Pizarro, and S. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1027–1034.
- [44] M. Damghanian, R. Olsson, M. Sjöström, A. Erdmann, and C. Perwaß, "Spatial resolution in a multi-focus plenoptic camera," in *International Conference on Image Processing (ICIP)*, 2014.
- [45] L. Palmieri and R. Koch, "Optimizing the lens selection process for multi-focus plenoptic cameras and numerical evaluation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1763–1774.
- [46] Raytrix GmbH, "Raytrix RxLive Software," <https://raytrix.de/downloads/>, accessed on 07.03.2019.
- [47] R. Ferreira and N. Gonçalves, "Fast and accurate micro lenses depth maps for multi-focus light field cameras," in *German Conference on Pattern Recognition (GCPR)*, 2016.
- [48] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand *et al.*, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.
- [49] A. Waqas, S. Märten, and O. Roger, "Compression scheme for sparsely sampled light field data based on pseudo multi-view sequences," in *Optics, Photonics, and Digital Technologies for Imaging Applications V*, vol. 10679. International Society for Optics and Photonics, 2018, p. 106790M.
- [50] R. Ferreira and N. Gonçalves, "Accurate and fast micro lenses depth maps from a 3d point cloud in light field cameras," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 1893–1898.
- [51] A. Ghasemi, N. Afonso, and M. Vetterli, "Lcav-31: A dataset for light field object recognition," in *Proceedings of the SPIE, vol. 9020, International Society for Optics and Photonics*, 2014.
- [52] C. Hazirbas, "4.5d lightfield-depth benchmark," <http://hazirbas.com/datasets/ddff12scene/>, [Online, 2018].
- [53] A. Mousnier, E. Vural, and C. Guillemot, "Lytro first generation dataset," <https://www.irisa.fr/temics/demos/lightField/index.html>, [Online, 2018].
- [54] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, no. EPPL-CONF-218363, 2016.
- [55] K. Honauer and O. J. *et al.*, "A taxonomy and evaluation of dense light field depth estimation algorithms," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [56] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4d light fields," in *VMV*. Citeseer, 2013, pp. 225–226.
- [57] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 43–54.

Bibliography

- [AB91] 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.
- [AP] Waqas Ahmad and Luca Palmieri. *Matching light field datasets from plenoptic cameras 1.0 and 2.0*. <https://doi.org/10.6084/m9.figshare.6115487>. [Online, 2020].
- [APK+18] Waqas Ahmad, Luca Palmieri, Reinhard Koch, and Märten Sjöström. “Matching light field datasets from plenoptic cameras 1.0 and 2.0”. In: *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE. 2018, pp. 1–4.
- [AW92] Edward H Adelson and John Y. A. Wang. “Single lens stereo with a plenoptic camera”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 2 (1992), pp. 99–106.
- [BBM87] Robert C Bolles, H Harlyn Baker, and David H Marimont. “Epipolar-plane image analysis: an approach to determining structure from motion”. In: *International Journal of Computer Vision* 1.1 (1987), pp. 7–55.
- [BGY+13] Michael Broxton, Logan Grosenick, Samuel Yang, Noy Cohen, Aaron Andalman, Karl Deisseroth, and Marc Levoy. “Wave optics theory and 3-d deconvolution for the light field microscope”. In: *Optics express* 21.21 (2013), pp. 25418–25439.
- [BHK16] Yunsu Bok, Hyowon Ha, and In So Kweon. “Automated checkerboard detection and indexing using circular boundaries”. In: *Pattern Recognition Letters* 71 (2016), pp. 66–72.

Bibliography

- [BJK17] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. “Geometric calibration of micro-lens-based light field cameras using line features”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.2 (2017), pp. 287–300.
- [Ble20] Blender. *Open source 3d creation suite*. <https://www.blender.org>. [Online, Accessed on July, 2020].
- [BSH+17] Noah Bedard, Timothy Shope, Alejandro Hoberman, Mary Ann Haralam, Nader Shaikh, Jelena Kovačević, Nikhil Balram, and Ivana Tošić. “Light field otoscope design for 3d in vivo imaging of the middle ear”. In: *Biomed. Opt. Express* 8.1 (Jan. 2017), pp. 260–272. DOI: 10.1364/BOE.8.000260. URL: <http://www.osapublishing.org/boe/abstract.cfm?URI=boe-8-1-260>.
- [CLS16] Hao Chen, Peter M Lillo, and Volker Sick. “Three-dimensional spray–flow interaction in a spark-ignition direct-injection engine”. In: *International Journal of Engine Research* 17.1 (2016), pp. 129–138. DOI: 10.1177/1468087415608741.
- [CLY+14] Can Chen, Haiting Lin, Zhan Yu, Sing Bing Kang, and Jingyi Yu. “Light field stereo matching using bilateral statistics of surface cameras”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1518–1525.
- [CYA+14] Noy Cohen, Samuel Yang, Aaron Andalman, Michael Broxton, Logan Grosenick, Karl Deisseroth, Mark Horowitz, and Marc Levoy. “Enhancing the performance of the light field microscope using wavefront coding”. In: *Optics express* 22.20 (2014), pp. 24817–24839.
- [doi20] doitplenoptic. *Light field microscopy*. <https://www.doitplenoptic.com/>. [Online, Accessed on September, 2020].
- [DOS+14] M. Damghanian, R. Olsson, M. Sjöström, A. Erdmann, and C. Perwaß. “Spatial resolution in a multi-focus plenoptic camera”. In: *International Conference on Image Processing (ICIP)*. 2014.

- [DPW13] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams. “Decoding, calibration and rectification for lenselet-based plenoptic cameras”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 1027–1034.
- [DSF+17] Donald G Dansereau, Glenn Schuster, Joseph Ford, and Gordon Wetzstein. “A wide-field-of-view monocentric light field camera”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5048–5057.
- [FG16] Rodrigo Ferreira and Nuno Goncalves. “Fast and accurate micro lenses depth maps for multi-focus light field cameras”. In: *German Conference on Pattern Recognition*. Springer. 2016, pp. 309–319.
- [FK14] Oliver Fleischmann and Reinhard Koch. “Lens-based depth estimation for multi-focus plenoptic cameras”. In: *German Conference on Pattern Recognition*. Springer. 2014, pp. 410–420.
- [Gab08] Lippmann Gabriel. “La photographie intégrale”. In: *Comptes-Rendus, Académie des Sciences* 146 (1908), pp. 446–551.
- [GGS+96] 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*. ACM. 1996, pp. 43–54.
- [GL10] Todor Georgiev and Andrew Lumsdaine. “Reducing plenoptic camera artifacts”. In: *Comput. Graph. Forum* 29 (2010), pp. 1955–1968.
- [GL12] Todor Georgiev and Andrew Lumsdaine. “The multifocus plenoptic camera”. In: *Digital Photography VIII*. Vol. 8299. International Society for Optics and Photonics. 2012, p. 829908.
- [Goo20] Google. *About google*. <https://about.google/>. [Februar, 2020].
- [GZC+06] Todor Georgiev, Ke Colin Zheng, Brian Curless, David Salesin, Shree Nayar, and Chintan Intwala. “Spatio-angular resolution tradeoffs in integral photography.” In: *Rendering Techniques 2006* (2006), pp. 263–272.

Bibliography

- [HAH+14a] Christopher Hahne, Amar Aggoun, Shyqyri Haxha, Vladan Velisavljevic, and Juan Carlos Jácome Fernández. “Light field geometry of a standard plenoptic camera”. In: *Optics express* 22.22 (2014), pp. 26659–26673.
- [HAH+14b] Christopher Hahne, Amar Aggoun, Shyqyri Haxha, Vladan Velisavljevic, and Juan CJ Fernández. “Baseline of virtual cameras acquired by a standard plenoptic camera setup”. In: *2014 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE. 2014, pp. 1–3.
- [HAV+16] Christopher Hahne, Amar Aggoun, Vladan Velisavljevic, Susanne Fiebig, and Matthias Pesch. “Refocusing distance of a standard plenoptic camera”. In: *Optics Express* 24.19 (2016), pp. 21521–21540.
- [Hec98] E. Hecht. *Optics*. 4th. Addison-Wesley, 1998.
- [Hir07] Heiko Hirschmuller. “Stereo processing by semiglobal matching and mutual information”. In: *IEEE Transactions on pattern analysis and machine intelligence* 30.2 (2007), pp. 328–341.
- [HJK+16] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. “A dataset and evaluation methodology for depth estimation on 4d light fields”. In: *Asian Conference on Computer Vision*. Springer. 2016, pp. 19–34.
- [HS08] Heiko Hirschmuller and Daniel Scharstein. “Evaluation of stereo matching costs on images with radiometric differences”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.9 (2008), pp. 1582–1599.
- [HSH+16] Christian Heinze, Stefano Spyropoulos, Stephan Hussmann, and Christian Perwass. “Automated robust metric calibration algorithm for multifocus plenoptic cameras”. In: *IEEE Transactions on Instrumentation and Measurement* 65.5 (2016), pp. 1197–1205.

- [HSJ+14] Matthew Hirsch, Sriram Sivaramakrishnan, Suren Jayasuriya, Albert Wang, Alyosha Molnar, Ramesh Raskar, and Gordon Wetzstein. “A switchable light field camera architecture with angle sensitive pixels and dictionary-based sparse coding”. In: *2014 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2014, pp. 1–10.
- [HSV+17] Matthieu Hog, Neus Sabater, Benoît Vandame, and Valter Drazic. “An image rendering pipeline for focused plenoptic cameras”. In: *IEEE Transactions on Computational Imaging* 3.4 (2017), pp. 811–821.
- [IKT+18] Yasutaka Inagaki, Yuto Kobayashi, Keita Takahashi, Toshiaki Fujii, and Hajime Nagahara. “Learning to capture light fields through a coded aperture camera”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 418–434.
- [JHG+13] O. Johannsen, C. Heinze, B. Goldluecke, and C. Perwaß. “On the calibration of focused plenoptic cameras”. In: *Proc. Seminar Time-Flight Depth Imag. Sens. Algorithms Appl.* 2013.
- [JPC+15] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. “Accurate depth map estimation from a lenslet light field camera”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1547–1555.
- [KLK+18] Min Seok Kim, Gil Ju Lee, Hyun Myung Kim, Hyuk Jae Jang, and Young Min Song. “Mobile terminal and control method for the mobile terminal”. U.S. pat. 10135963. LG Electronics. Nov. 20, 2018. URL: <https://patents.justia.com/patent/10135963>.
- [KLK+19] Min Seok Kim, Gil Ju Lee, Hyun Myung Kim, Hyuk Jae Jang, and Young Min Song. “Light field imaging with a hand-held smartphone camera for portable augmented reality applications”. In: *Imaging and Applied Optics 2019 (COSI, IS, MATH, pcAOP)*. Optical Society of America, 2019, ITu2C.3. DOI: 10.1364/ISA.2019.ITu2C.3. URL: <http://www.osapublishing.org/abstract.cfm?URI=ISA-2019-ITu2C.3>.

Bibliography

- [LCB+15] Haiting Lin, Can Chen, Sing Bing Kang, and Jingyi Yu. “Depth recovery from light field using focal stack symmetry”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3451–3459.
- [LFD+07] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. “Image and depth from a conventional camera with a coded aperture”. In: *ACM transactions on graphics (TOG)* 26.3 (2007), 70–es.
- [LFD08] Anat Levin, William T Freeman, and Frédo Durand. “Understanding camera trade-offs through a bayesian analysis of light field projections”. In: *European Conference on Computer Vision*. Springer. 2008, pp. 88–101.
- [LG09] Andrew Lumsdaine and Todor Georgiev. “The focused plenoptic camera”. In: *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE. 2009, pp. 1–8.
- [LH96] Marc Levoy and Pat Hanrahan. “Light field rendering”. In: *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM. 1996, pp. 31–42.
- [Lig20] Light. *Light technology*. <https://light.co/technology>. [Februar, 2020].
- [LLU+18] Huimin Lu, Yujie Li, Tomoki Uemura, Hyoungeop Kim, and Seiichi Serikawa. “Low illumination underwater light field images reconstruction using deep convolutional neural networks”. In: *Future Generation Computer Systems* 82 (2018), pp. 142–148.
- [LNA+06] Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. “Light field microscopy”. In: *ACM Transactions on Graphics (TOG)*. Vol. 25. 3. ACM. 2006, pp. 924–934.
- [LND17] Chao Liu, Srinivasa G Narasimhan, and Artur W Dubrawski. “Matting and depth recovery of thin structures using a focal stack”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6970–6978.

- [Lyt18] Lytro. *Home*. <https://www.lytro.com>. [Not Available, Accessed on March, 2018].
- [LZM09] Marc Levoy, Zhengyun Zhang, and Ian McDowall. "Recording and controlling the 4d light field in a microscope using microlens arrays". In: *Journal of microscopy* 235.2 (2009), pp. 144–162.
- [MBG18] Nuno Barroso Monteiro, Joao Pedro Barreto, and José Gaspar. "Surface cameras from shearing for disparity estimation on a lightfield". In: *RECPAD - Portuguese Conference on Pattern Recognition*. 2018.
- [MI15] Lois Mignard-Debise and Ivo Ihrke. "Light-field microscopy with a consumer light-field camera". In: *3D Vision (3DV), 2015 International Conference on*. IEEE. 2015, pp. 335–343.
- [MKR+17] Zak Murez, David Kriegman, Ravi Ramamoorthi, et al. "Depth and image restoration from light field in a scattering medium". In: (2017).
- [MPP+] Tim Michels, Arne Petersen, Luca Palmieri, and Reinhard Koch. *Camera generator add-on for blender 2.8*. <https://github.com/Arne-Petersen/Plenoptic-Simulation>. [Online, 2020].
- [MPP+18] T. Michels, A. Petersen, L. Palmieri, and R. Koch. "Simulation of plenoptic cameras". In: *2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. June 2018, pp. 1–4. DOI: 10.1109/3DTV.2018.8478432.
- [MRI17] L. Mignard-Debise, J. Restrepo, and I. Ihrke. "A unifying first-order model for light-field cameras: the equivalent camera array". In: *IEEE Transactions on Computational Imaging* 3.4 (2017), pp. 798–810.
- [MUG18] Ehsan Miandji, Jonas Unger, and Christine Guillemot. "Multi-shot single sensor light field camera using a color coded mask". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 226–230.

Bibliography

- [MWB+13] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. “Compressive Light Field Photography using Overcomplete Dictionaries and Optimized Projections”. In: *ACM Trans. Graph. (Proc. SIGGRAPH)* 32.4 (2013), pp. 1–11.
- [NCP+17] Sotiris Nousias, Francois Chadebecq, Jonas Pichat, Pearse Keane, Sebastien Ourselin, and Christos Bergeles. “Corner-based geometric calibration of multi-focus plenoptic cameras”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 957–965.
- [NLM+05] R. Ng, M. Levoy, B. Mathieu, G. Duval, M. Horowitz, and P. Hanrahan. “Light field photography with a hand-held plenoptic camera”. In: *Computer Science Technical Report CSTR 2.11* (2005), pp. 1–11.
- [Nok20] Nokia. *Nokia 9 pureview*. https://www.nokia.com/phones/en_us/nokia-9-pureview. [Februar, 2020].
- [OEE+18] Ryan Styles Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. “A system for acquiring, compressing, and rendering panoramic light field stills for virtual reality”. In: *ACM Trans. Graph.* 37 (2018).
- [Pala] Luca Palmieri. *Multi-focus plenoptic images dataset*. <https://data.mendeley.com/datasets/t6czryg5nw/draft?a=88c67f62-d5de-4c40-87a9-cdaad3141081>. [Online, 2018].
- [Palb] Luca Palmieri. *Robust depth estimation for light field microscopy*. <https://github.com/PlenopticToolbox/RobustDepthLFMicroscopy>. [Online, 2020].
- [Pal20] Luca Palmieri. *The plenoptic toolbox 2.0*. <https://github.com/PlenopticToolbox/PlenopticToolbox2.0>. [Online, 2020].
- [PK17] Luca Palmieri and Reinhard Koch. “Optimizing the lens selection process for multi-focus plenoptic cameras and numerical evaluation”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1763–1774.

- [PKV18] 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)*. IEEE. 2018, pp. 649–653.
- [PPG13] Said Pertuz, Domenech Puig, and Miguel Angel Garcia. “Analysis of focus measure operators for shape-from-focus”. In: *Pattern Recognition* 46.5 (2013), pp. 1415–1432.
- [PSI+19] Luca Palmieri, Gabriele Scrofani, Nicolò Incardona, Genaro Saavedra, Manuel Martínez-Corral, and Reinhard Koch. “Robust depth estimation for light field microscopy”. In: *Sensors* 19.3 (2019), p. 500.
- [PW12] Christian Perwass and Lennart Wietzke. “Single lens 3d-camera with extended depth-of-field”. In: *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics. 2012, pp. 829108–829108.
- [Ray20] Raytrix. *Light field technology*. <https://www.raytrix.de>. [Online, Accessed on February, 2020].
- [SAG17] Michael Strecke, Anna Alperovich, and Bastian Goldluecke. “Accurate depth and normal maps from occlusion-aware focal stack symmetry”. In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE. 2017, pp. 2529–2537.
- [SB17] David C Schedl and Oliver Bimber. “Compressive volumetric light-field excitation”. In: *Scientific reports* 7.1 (2017), pp. 1–9.
- [SJY+18] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. “Epinet: a fully-convolutional neural network using epipolar geometry for depth from light field images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4748–4757.

Bibliography

- [SPS+19] Anca Stefanoiu, Josue Page, Panagiotis Symvoulidis, Gil G Westmeyer, and Tobias Lasser. “Artifact-free deconvolution in light field microscopy”. In: *Optics express* 27.22 (2019), pp. 31644–31666.
- [SSL+18] G Scrofani, J Sola-Pikabea, A Llavador, Emilio Sanchez-Ortiga, JC Barreiro, G Saavedra, J Garcia-Sucerquia, and Manuel Martínez-Corral. “Fimic: design for ultimate 3d-integral microscopy of in-vivo biological samples”. In: *Biomedical optics express* 9.1 (2018), pp. 335–346.
- [SWG+11] Sriram Sivaramakrishnan, Albert Wang, Patrick R Gill, and Alyosha Molnar. “Enhanced angle sensitive pixels for light field imaging”. In: *2011 International Electron Devices Meeting*. IEEE. 2011, pp. 8–6.
- [THM+13] Michael W Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. “Depth from combining defocus and correspondence using light-field cameras”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 673–680.
- [TSM+15] Michael W Tao, Pratul P Srinivasan, Jitendra Malik, Szymon Rusinkiewicz, and Ravi Ramamoorthi. “Depth from shading, defocus, and correspondence using light-field angular coherence”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1940–1948.
- [VAD+18] Alessandro Vianello, Jens Ackermann, Maximilian Diebold, and Bernd Jähne. “Robust hough transform based 3d reconstruction from circular light fields”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.
- [VRA+07] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. “Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing”. In: *ACM transactions on graphics (TOG)*. Vol. 26. 3. ACM. 2007, p. 69.

- [WER15] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. "Occlusion-aware depth estimation using light-field cameras". In: *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE. 2015, pp. 3487–3495.
- [WFJ11] Sven Wanner, Janis Fehr, and Bernd Jähne. "Generating epi representations of 4d light fields with a single lens focused plenoptic camera". In: *International Symposium on Visual Computing*. Springer. 2011, pp. 90–101.
- [WG12] Sven Wanner and Bastian Goldluecke. "Globally consistent depth labeling of 4d light fields". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 41–48.
- [WMG13] Sven Wanner, Stephan Meister, and Bastian Goldluecke. "Datasets and benchmarks for densely sampled 4d light fields." In: *VMV*. Citeseer. 2013, pp. 225–226.
- [WSG13] Sven Wanner, Christoph Straehle, and Bastian Goldluecke. "Globally consistent multi-label assignment on the ray space of 4d light fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 1011–1018.
- [WZH+16] Ting-Chun Wang, Jun-Yan Zhu, Ebi Hiroaki, Manmohan Chandraker, Alexei A Efros, and Ravi Ramamoorthi. "A 4d light-field dataset and cnn architectures for material recognition". In: *European Conference on Computer Vision*. Springer. 2016, pp. 121–138.
- [YGL+13] Zhan Yu, Xinqing Guo, Haibing Lin, Andrew Lumsdaine, and Jingyi Yu. "Line assisted light field triangulation and stereo matching". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2792–2799.
- [ZSL+16] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. "Robust depth estimation for light field via spinning parallelogram operator". In: *Computer Vision and Image Understanding* 145 (2016), pp. 148–159.

Bibliography

- [ZZL+18] Qi Zhang, Chunping Zhang, Jinbo Ling, Qing Wang, and Jingyi Yu. "A generic multi-projection-center model and calibration method for light field cameras". In: *IEEE transactions on pattern analysis and machine intelligence* (2018).