From the Calibration of a Light-Field Camera to Direct Plenoptic Odometry

Niclas Zeller, Franz Quint, and Uwe Stilla

***Abstract*—This paper presents a complete framework from the calibration of a plenoptic camera toward plenoptic camera based visual odometry. This is achieved by establishing the multiple view geometry for plenoptic cameras. Based on this novel multiple view geometry, a calibration approach is developed. The approach optimizes all intrinsic parameters of the plenoptic camera model, the 3D coordinates of the calibration points, and all camera poses in a single bundle adjustment. Our plenoptic camera based visual odometry algorithm, called direct plenoptic odometry (DPO), is a direct and semi-dense approach, which takes advantage of the full sensor resolution. DPO also relies on our multiple view geometry for plenoptic cameras. Tracking and mapping works directly on the micro images formed by the micro lens array and, therefore, has not to deal with aliasing effects in the spatial domain. The algorithm generates a semi-dense depth map based on correspondences between subsequent light-field frames, while taking differently focused micro images into account. Up to our knowledge, it is the first method that performs tracking and mapping for plenoptic cameras directly on the micro images. DPO outperforms state-of-the-art direct monocular simultaneous localization and mapping (SLAM) algorithms and can compete in accuracy with latest stereo SLAM approaches, while supplying much more detailed point clouds.**

***Index Terms*—Light-field, plenoptic camera calibration, plenoptic multiple view geometry, SLAM, visual odometry.**

I. Introduction

VISUAL Odometry (VO) and Simultaneous Localization and Mapping (SLAM) currently are some of the most addressed tasks in computer vision. Such approaches allow for tracking and mapping in unknown environments without any further infrastructure (e.g., GPS). In the last years VO and SLAM approaches for monocular and stereo cameras as well as active RGB-D sensors (e.g., structured light sensors) were driven forward. Nevertheless, while pure monocular approaches lack from scale awareness, stereo cameras or RGB-D sensors in general have large dimensions and therefore are impractical for certain applications.

During the last decades plenoptic cameras gained more and more interest [1]. While such cameras have dimensions similar to monocular cameras they are able to retrieve depth from single images based on the recorded light-field. Hence, plenoptic camera based VO would combine key features from both monocular and stereo approaches and closes the gap between the two.

In this article we present the complete workflow for plenoptic camera based VO. We introduce a new mathematical model for micro lens array (MLA) based light-field cameras and define a plenoptic multiple view geometry based on this model. This multiple view geometry leads us to a plenoptic camera calibration approach and builds the foundation for the Direct Plenoptic Odometry (DPO) algorithm. DPO combines advantages of monocular VO, like a single sensor system or scale invariance with static stereo and scale awareness at least for object distances in the range of a few meters. Our algorithm creates highly detailed, semi-dense point clouds (see Fig. 1) by working directly on pixel intensities in the micro images of a plenoptic camera.

A. Related Work

1) Plenoptic Camera Calibration: During the last years different methods were published to calibrate plenoptic cameras.

A method for correcting aberrations of the main lens based on the recorded 4D light-field inside the camera is presented in [2]. [3] presents a complete pipeline for the calibration of the MLA based on the specific case of a Lytro camera.

A complete mathematical model for unfocused plenoptic cameras is derived for the first time in [4]. To overcome the problem of feature point detection in the small micro images, [5] presents a calibration method which makes use of line features extracted from the micro images.

There exist several methods to perform calibration of focused plenoptic cameras [6]–[10]. All these methods model the projection from the object space to the virtual image space instead of projecting directly to the micro images. While [6] and [7] rely on a planar calibration target, [10] estimates the camera parameters based on a 3D object.

Manuscript received November 30, 2016; revised April 16, 2017 and June 22, 2017; accepted July 12, 2017. Date of publication August 11, 2017; date of current version October 23, 2017. This work was supported in part by the Baden-Württemberg-Stiftung in its program "Photonics, Microelectronics, Information Technology" and in part by the Federal Ministry of Education and Research of Germany in its program "FHprofUnt." The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Qing Wang. *(Corresponding author: Niclas Zeller.)*

N. Zeller is with the Faculty of Electrical Engineering and Information Technology, Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany, and also with Technische Universität München, 80333 München, Germany (e-mail: niclas.zeller@hs-karlsruhe.de).

F. Quint is with the Faculty of Electrical Engineering and Information Technology, Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany (e-mail: franz.quint@hs-karlsruhe.de).

U. Stilla is with the Department of Photogrammetry and Remote Sensing, Technische Universität München, 80333 München, Germany (e-mail: stilla@tum.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2737965

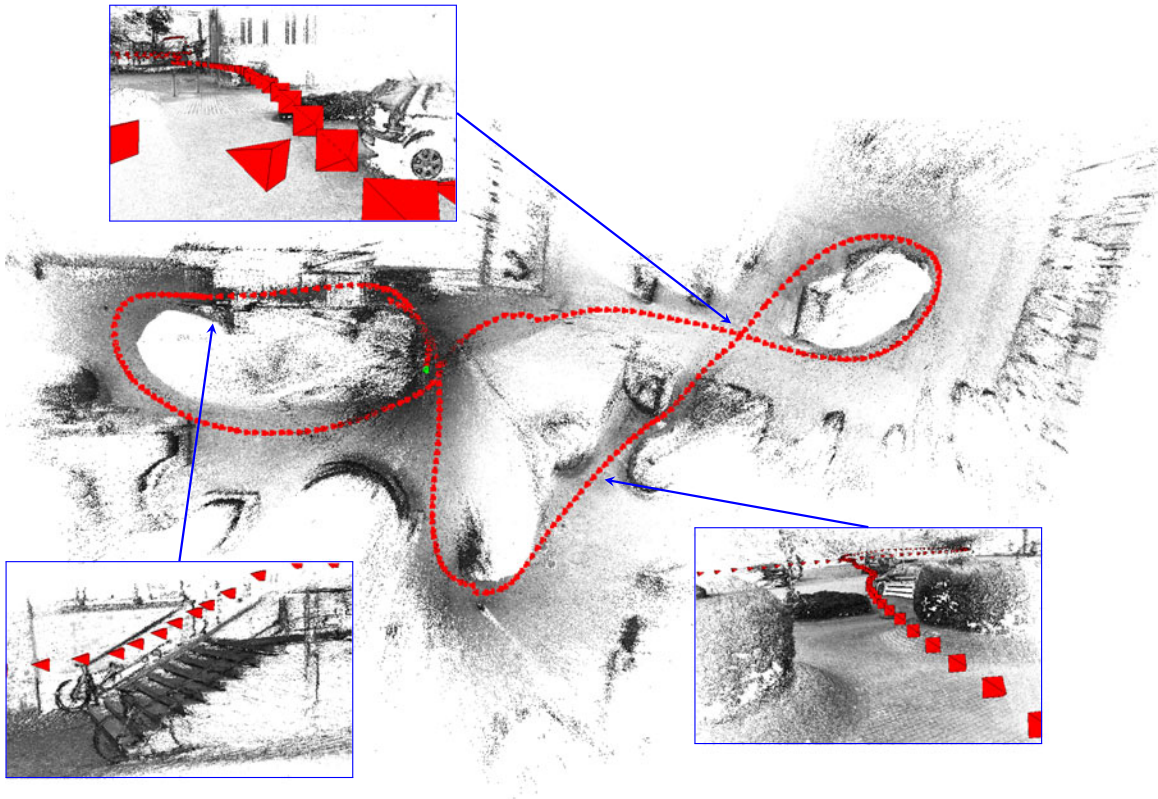


Fig. 1. Semi-dense point cloud of "parking lot" sequence created by Direct Plenoptic Odometry (DPO). The figure shows the top view of the complete trajectory and detailed subsections of the point cloud. Keyframe positions are marked in red, while the green dot represents the end of the trajectory.

2) Visual Odometry and SLAM: Feature based SLAM reduces the amount of data and therefore the computational effort by extracting a set of meaningful feature points from the recorded images. Afterwards, camera orientation and scene structure are estimated only based on this sparse set of geometric feature points [11]–[15].

Direct methods, like the one presented in [16], perform tracking and mapping directly on the recorded images. Tracking becomes much more robust since all image data is used. Here, tracking and dense mapping can be performed in one step, as shown in [17]. The complexity of such algorithms can be reduced by considering only textured image regions [18], [19]. These semi-dense, direct methods are capable to run in real-time on today's standard CPUs or even on smartphones [20].

While for monocular approaches camera motion is needed to obtain scene structure, the SLAM problem significantly simplifies when using stereo or RGB-D cameras. Since in this case absolute depth is received from a static recording, such approaches are able to measure the scale of the scene without using any additional sensors [21]–[24].

There exist few VO methods based on light-field representation [25], [26]. Anyhow, the cited methods mainly have been developed for light-fields recorded by camera arrays with large stereo baselines. In [27] a feature based Structure from Motion (SfM) approach working on 4D light-field representations is presented.

B. Contribution of This Work

This article presents a complete framework for plenoptic VO. We establish the multiple view geometry for MLA based light-field cameras. By working directly on the recorded micro images (raw image) we are able to find stereo correspondences using full sensor resolution instead of low resolution sub-aperture images. This avoids aliasing effects due to undersampling in the spatial domain [28] and results in much higher resolved depth maps.

Furthermore, we show that full resolution stereo matching leads to larger effective stereo baselines in comparison to the low resolution sub-aperture images.

We present a calibration approach which performs a complete bundle adjustment. Here, all intrinsic camera parameters, the 3D object coordinates and all camera poses are estimated in a single optimization task. The calibration is based on a 3D calibration target, which, as will be shown, significantly improves the calibration result compared to using a planar calibration target.

A major contribution is also the semi-dense direct visual odometry algorithm for MLA based light-field cameras, which makes use of the introduced multiple view geometry and the estimated camera model. We define a focus disparity error which models the effect of differently focused micro images on the depth estimation.

To the best of our knowledge, this is the first visual odometry approach which works directly on the recorded raw data of a

plenoptic camera and improves the static depth by plenoptic SfM.

C. Outline

In Section II we give some preliminary definitions used throughout this paper. Section III gives a brief, theoretical background on plenoptic cameras. The following methodical part consists of three main sections. Section IV defines the multiple view geometry for plenoptic cameras. The bundle adjustment based calibration approach is presented in Section V and Section VI describes the DPO algorithm. Finally, Section VII presents the evaluation of our methods, while Section VIII summarizes and concludes this work.

II. Preliminaries

Throughout the paper we use several notations, definitions and symbols. To enhance the readability of the paper, important notations, definitions and symbols are introduced here.

We denote matrices as bold, capital letters ($\boldsymbol{G}$), vectors as bold, lower case letters ($\boldsymbol{x}$) and scalars as normal letters, either capital or lower case (d). We do not differentiate between homogeneous ($\boldsymbol{x} = (x, y, z, 1)^T$) and non-homogeneous ($\boldsymbol{x} = (x, y, z)^T$) representations. Nevertheless, this should be clear from the context.

Throughout the paper rigid body transformations are defined based on Lie-Manifolds. Hence, a rigid body transformation $\boldsymbol{G} \in \mathrm{SE}(3)$ is completely defined by the six-dimensional vector $\boldsymbol{\xi} \in \mathbb{R}^6$, which defines an element of the corresponding Lie-Algebra $\mathfrak{se}(3)$. For more details on Lie-Manifolds we refer to [29].

The notation $[\boldsymbol{t}]_i$ defines the i-th element of a vector $\boldsymbol{t}$ and $[\boldsymbol{R}]_j$ the j-th row of a matrix $\boldsymbol{R}$.

The symbol δ_{mn} defines a Kronecker delta as given in eq. (1).

$$\delta_{mn} = \begin{cases} 1 & \text{for } m = n, \\ 0 & \text{for } m \neq n. \end{cases} \tag{1}$$

For the plenoptic camera we differentiate between image space and object space. To receive an upright image in image coordinates of an upright object in object coordinates, we define image coordinates to be mirrored with respect to object coordinates. Thus, to transform from object to image coordinates one has to take the negative coefficients of the respective vector.

Table I shows all important symbols used in this paper.

TABLE I
List of Important Symbols

symbol	description
v	virtual depth
b_L	image distance
$b_{L\,0}$	distance main lens to MLA
B	distance MLA to sensor
f_L	main lens focal length
f_M	micro lens focal length
D_M	micro lens diameter
$\boldsymbol{c}_L$	main lens principal point
$c_{L\,x}$	main lens principal point x-coord.
$c_{L\,y}$	main lens principal point y-coord.
$\boldsymbol{c}_{M\,L}$	micro lens center
$\boldsymbol{p}_{M\,L}$	projected micro lens center
$z_{C\,0}$	effective object distance offset
z_C	object distance
$z'_C = \lambda$	effective object distance
μ	pixel disparity
μ_p	projected micro image disparity
$\boldsymbol{x}_C$	camera coordinates
$\boldsymbol{x}'_C$	effective camera coordinates
$\boldsymbol{x}_R$	raw image coordinates
$\boldsymbol{x}_{M\,L}$	micro image coordinates
$\boldsymbol{x}_p$	projected micro image coordinates
$\boldsymbol{R}$	rotation matrix $\in \mathrm{SO}(3)$
$\boldsymbol{t}$	translation vector $\in \mathbb{R}^3$
$\boldsymbol{t}'$	effective translation vector $\in \mathbb{R}^3$
$\boldsymbol{\xi}$	element of the Lie-algebra $\mathfrak{se}(3)$
$\boldsymbol{G}$	rigid body transformation $\in \mathrm{SE}(3)$
$\boldsymbol{G}'$	effective rigid body trans. $\in \mathrm{SE}(3)$
d	inverse effective depth $d = \lambda^{-1}$
σ_d^2	inverse effective depth variance
$D_{M\,L}(\boldsymbol{x}_R)$	prob. micro image depth map
$D_V(\boldsymbol{x}_V)$	prob. virtual image depth map
$I_{M\,L}(\boldsymbol{x}_R)$	raw image (micro images)
$I_V(\boldsymbol{x}_V)$	virtual intensity image
$\sigma^2_{\mu(\xi,\pi)}$	variance of geometric disp. error
$\sigma^2_{\mu(I)}$	variance of photometric disp. error
$\sigma^2_{\mu(v,k,j)}$	variance of focus disparity error
$\pi_{M\,L}(\cdot)$	camera projection on micro images
$\pi_V(\cdot)$	camera projection on virtual image
$\delta_{m\,n}$	Kronecker delta

III. The Plenoptic Camera

Placing a MLA in front of the image sensor transforms a regular camera into a plenoptic camera, which is able to record 4D light-field information in a single image. While all plenoptic cameras rely on this concept, such cameras can be realized in different configurations depending on the camera parameters. In general we separate between unfocused plenoptic cameras (plenoptic camera 1.0) [30], [31] and focused plenoptic cameras (plenoptic camera 2.0) [32].

Focused plenoptic cameras produce focused micro images on the sensor, where each micro image captures a small portion of the complete scene. Due to the redundancy in the micro images, stereo correspondences can be found directly in the micro images using full sensor resolution. Hence, prior to any metric calibration one is able to estimate a so called *virtual depth* $v = \frac{b_L - b_{L\,0}}{B}$, which is a measure for the image distance b_L of the *virtual* main lens *image* [9] (see Fig. 2). Based on the virtual depth v one is able to project points from the focused micro images to the virtual image and thereby reconstruct the corresponding intensity image. Since this reconstructed image has a very large depth of field (DOF), we will call it the *totally focused image*.

From the recordings of unfocused plenoptic cameras one is able to extract a set of so called *sub-aperture images*. All of these sub-aperture images show the complete recorded scene from slightly different perspectives. The resolution of a sub-aperture image is limited by the number of micro lenses in the

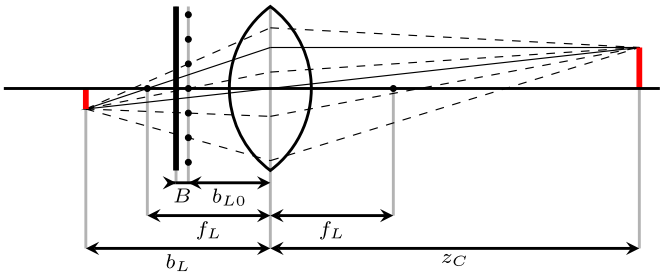


Fig. 2. Projection process of a focused plenoptic camera. MLA in distance b_{L0} to the main lens produces focused micro images of the virtual main lens image on the sensor.

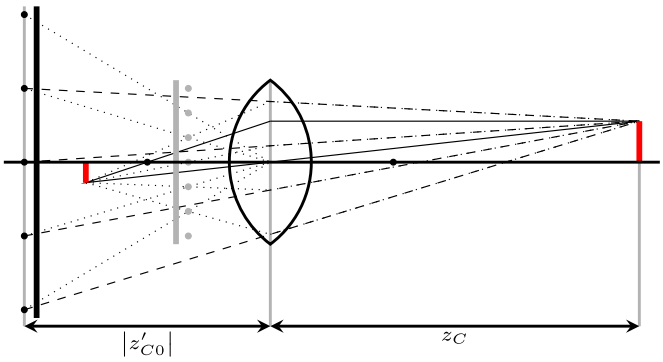


Fig. 3. Focused plenoptic camera interpreted as an array of very narrow FOV pinhole cameras at distance $|z'_{C0}|$ to the main lens.

MLA and therefore can not be increased by increasing the sensor resolution.

IV. Multiple View Stereo for Plenoptic Cameras

We show that a MLA based light-field camera can be interpreted as an array of very narrow field of view (FOV) pinhole cameras observing the scene. From this new interpretation insight into full resolution multiple view epipolar geometry for plenoptic cameras is obtained. Even though this interpretation is predestined for focused plenoptic cameras, such full resolution approaches can also be applied to unfocused plenoptic cameras, which supply partly focused micro images [33].

A. Plenoptic Camera Interpretation

Fig. 2 shows the projection process of a focused plenoptic camera. In the same way as the object points are projected through the main lens to corresponding image points, the micro lens centers can be projected from image space into object space based on the thin lens equation. Thus, the resulting object distance z'_{C0} of a micro lens center is defined as given in eq. (2).

$$z'_{C0} = \frac{f_L \cdot b_{L0}}{b_{L0} - f_L} \tag{2}$$

Here f_L is the focal length of the main lens and b_{L0} the distance between MLA and main lens plane. Fig. 3 shows the projected micro lens centers, resulting in a virtual camera array.

From Fig. 3 one can see that for the given setup ($f_L > b_{L0}$) the micro lenses are projected behind the main lens.

For later use we define z_{C0} as the negative value of z'_{C0}:

$$z_{C0} := -z'_{C0} = \frac{f_L \cdot b_{L0}}{f_L - b_{L0}} \tag{3}$$

With this definition, z'_C given in eq. (4) represents the object distance with respect to the virtual camera array.

$$z'_C := z_C + z_{C0} \tag{4}$$

The x- and y-coordinate of the center of a projected micro lens is received as the intersection between the projected MLA plane and the main lens' central ray through the corresponding real micro lens. Thus, one receives a projected micro lens center in camera coordinates $\boldsymbol{p}_{ML}$ from the coordinates of the real micro lens center $\boldsymbol{c}_{ML}$ as given in eq. (5).

$$\begin{aligned}\boldsymbol{p}_{ML} &= \begin{pmatrix} p_{MLx} \\ p_{MLy} \\ -z_{C0} \end{pmatrix} = -\boldsymbol{c}_{ML}\frac{z_{C0}}{b_{L0}} = -\begin{pmatrix} c_{MLx} \\ c_{MLy} \\ b_{L0} \end{pmatrix}\frac{z_{C0}}{b_{L0}} \\ &= -\boldsymbol{c}_{ML}\frac{f_L}{f_L - b_{L0}} = \boldsymbol{c}_{ML}\frac{f_L}{b_{L0} - f_L}\end{aligned} \tag{5}$$

Here, the minus in front of $\boldsymbol{c}_{ML}$ is due to the transformation from image coordinates to object coordinates.

The projected micro images are defined such that they have a normalized focal length and are parallel to the x-y-plane at $z_C = 1 - z_{C0}$. In this way the central projection from homogeneous 2D to 3D coordinates is just a scaling. A point $\boldsymbol{x}_p$ in the projected micro image is calculated based on a point $\boldsymbol{x}_{ML}$ in the real micro image as follows:

$$\begin{aligned}\boldsymbol{x}_p &= \begin{pmatrix} x_p \\ y_p \\ 1 \end{pmatrix} = \boldsymbol{x}_{ML}\frac{f_L - b_{L0}}{f_L \cdot B} + \frac{\boldsymbol{c}_{ML}}{f_L} \\ &= \begin{pmatrix} x_{ML} \\ y_{ML} \\ B \end{pmatrix}\frac{f_L - b_{L0}}{f_L \cdot B} + \frac{\boldsymbol{c}_{ML}}{f_L}\end{aligned} \tag{6}$$

Beside the regular camera coordinates $\boldsymbol{x}_C$ with its origin in the center of the main lens, we define for each micro lens, i.e., for each virtual camera, separate camera coordinates $\boldsymbol{x}'_C$. This is necessary since each micro lens has different center coordinates $\boldsymbol{p}_{ML}$ and thus the respective camera coordinate systems are also different. In the sequel we will call $\boldsymbol{x}'_C$ *effective camera coordinates*. The origin of the effective camera coordinates $\boldsymbol{x}'_C$ is the corresponding projected micro lens center $\boldsymbol{p}_{ML}$. Therefore, we receive the following relation:

$$\boldsymbol{x}_C := \boldsymbol{x}'_C + \boldsymbol{p}_{ML} = z'_C\boldsymbol{x}_p + \boldsymbol{p}_{ML} \tag{7}$$

Even though Figs. 2 and 3 show only a setup for $f_L > b_{L0}$ the presented projection is valid for almost any setup. The only setup for which no projection is possible is for $f_L = b_{L0}$. Here the micro lenses are projected to infinity. For $f_L < b_{L0}$ it could even be the case that the MLA is projected behind the recorded scene and effectively observes it from the backside.

B. Multiple View Epipolar Geometry

Based on the plenoptic camera interpretation (Section IV-A) we are able to define a multiple view epipolar geometry for plenoptic cameras.

Consider a set of images (views) taken with the plenoptic camera from different locations. Let i and j be two particular views of this set. Eq. (8) defines the transformation of a 3D object point with camera coordinates $\boldsymbol{x}_C^{(i)}$ in the i-th view to camera coordinates $\boldsymbol{x}_C^{(j)}$ in the j-th view based on the rigid body transformation $\boldsymbol{G}_{ij} \in \mathrm{SE}(3)$.

$$\boldsymbol{x}_C^{(j)} = \boldsymbol{G}_{ij} \cdot \boldsymbol{x}_C^{(i)} = \begin{pmatrix} \boldsymbol{R}_{ij} & \boldsymbol{t}_{ij} \\ \boldsymbol{0} & 1 \end{pmatrix} \cdot \boldsymbol{x}_C^{(i)} \tag{8}$$

Similarly to a point in regular camera coordinates, a point in effective camera coordinates $\boldsymbol{x}_C^{\prime(i,k)}$ of the k-th micro lens in the i-th view can be transformed into the effective camera coordinates $\boldsymbol{x}_C^{\prime(j,l)}$ of the l-th micro lens in the j-th view based on a rigid body transformation:

$$\boldsymbol{x}_C^{\prime(j,l)} = \boldsymbol{G}_{ij}^{\prime(kl)} \cdot \boldsymbol{x}_C^{\prime(i,k)} \tag{9}$$

with $\boldsymbol{G}_{ij}^{\prime(kl)} \in \mathrm{SE}(3)$ defined as follows:

$$\boldsymbol{G}_{ij}^{\prime(kl)} = \begin{pmatrix} \boldsymbol{R}_{ij} & \boldsymbol{t}_{ij}^{\prime(kl)} \\ \boldsymbol{0} & 1 \end{pmatrix} \tag{10}$$

$$\boldsymbol{t}_{ij}^{\prime(kl)} = \boldsymbol{t}_{ij} - \boldsymbol{p}_{ML}^{(l)} + \boldsymbol{R}_{ij} \cdot \boldsymbol{p}_{ML}^{(k)} \tag{11}$$

Here $\boldsymbol{p}_{ML}^{(k)}$ and $\boldsymbol{p}_{ML}^{(l)}$ are the coordinates of the respective projected micro lens centers as defined in eq. (5).

Based on this definition one can further derive the epipolar geometry between one micro image in the i-th view and another micro image in the j-th view. From projected image coordinates $\boldsymbol{x}_p^{(i,k)}$ (i-th view) and $\boldsymbol{x}_p^{(j,l)}$ (j-th view) one receives the effective camera coordinates $\boldsymbol{x}_C^{\prime(i,k)}$ and $\boldsymbol{x}_C^{\prime(j,l)}$ as follows:

$$\boldsymbol{x}_C^{\prime(i,k)} = \lambda_i \boldsymbol{x}_p^{(i,k)} \quad \text{with} \quad \lambda_i := z_C^{\prime(i,k)} \tag{12}$$

$$\boldsymbol{x}_C^{\prime(j,l)} = \lambda_j \boldsymbol{x}_p^{(j,l)} \quad \text{with} \quad \lambda_j := z_C^{\prime(j,l)} \tag{13}$$

In order to enhance readability we omit in the following all indices which do not lead to ambiguous definitions ($\boldsymbol{x}_C^{\prime(i)} := \boldsymbol{x}_C^{\prime(i,k)}$; $\boldsymbol{x}_C^{\prime(j)} := \boldsymbol{x}_C^{\prime(j,l)}$; $\boldsymbol{R} := \boldsymbol{R}_{ij}$; $\boldsymbol{t}' := \boldsymbol{t}_{ij}^{\prime(kl)}$).

Inserting eq. (12) and (13) into eq. (9) leads to the following relation:

$$\boldsymbol{x}_C^{\prime(j)} = \lambda_j \boldsymbol{x}_p^{(j)} = \lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t}' \tag{14}$$

where λ_j can be written as follows:

$$\lambda_j = \lambda_i \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)} + \left[\boldsymbol{t}'\right]_3 \tag{15}$$

After combining eq. (14) and eq. (15) we receive a linear function in λ_i, as given in eq. (16).

$$\lambda_i \left(\left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)}\right) \boldsymbol{x}_p^{(j)} + \left[\boldsymbol{t}'\right]_3 \cdot \boldsymbol{x}_p^{(j)} = \lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t}' \tag{16}$$

Therefore, a point on the epipolar line in the micro image of the j-th frame is defined as follows:

$$\begin{aligned} \boldsymbol{x}_p^{(j)} &= \frac{\lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t}'}{\lambda_i \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)} + \left[\boldsymbol{t}'\right]_3} \\ &= \frac{\lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t} - \boldsymbol{p}_{ML}^{(l)} + \boldsymbol{R} \cdot \boldsymbol{p}_{ML}^{(k)}}{\lambda_i \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)} + \left[\boldsymbol{t}\right]_3 + z_{C0} + \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{p}_{ML}^{(k)}} \end{aligned} \tag{17}$$

Using this epipolar geometry we are able to deal with stereo observations from different micro images of different frames.

C. Effective Object Distances Versus Virtual Depth

For the case that the considered micro lenses are within the same frame ($\boldsymbol{t} = \boldsymbol{0}$ and $\boldsymbol{R} = (\delta_{mn})_{m,n \in \{1,2,3\}}$) eq. (17) simplifies as follows:

$$\boldsymbol{x}_p^{(j)} = \lambda_i^{-1} \left(\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\right) + \boldsymbol{x}_p^{(i)} \tag{18}$$

From eq. (18) we receive the disparity in the projected micro images μ_p as given in eq. (19).

$$\mu_p := \frac{\langle \boldsymbol{x}_p^{(j)} - \boldsymbol{x}_p^{(i)}, \boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)} \rangle}{\|\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\|} = \lambda_i^{-1} \|\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\| \tag{19}$$

From eq. (19) we see that for in-frame depth estimation the disparity μ_p is proportional to the inverse effective object distance $\lambda^{-1} = z_C^{\prime -1}$. In another paper [34] we have shown that the disparity μ in the real micro images is proportional to the inverse virtual depth v^{-1}. Since we performed a linear mapping from μ to μ_p, this leads us to the conclusion that there has to be a linear relationship between $z_C^{\prime -1}$ and v^{-1}. This is proven in the following:

$$\begin{aligned} z_C' &= z_C + z_{C0} = \left(\frac{1}{f_L} - \frac{1}{b_L}\right)^{-1} + z_{C0} \\ &= \left(\frac{1}{f_L} - \frac{1}{v \cdot B + b_{L0}}\right)^{-1} + \frac{f_L \cdot b_{L0}}{f_L - b_{L0}} \end{aligned} \tag{20}$$

Rearranging eq. (20) gives the following definition of $z_C^{\prime -1}$ with respect to v^{-1}:

$$z_C^{\prime -1} = -\frac{\left(f_L - b_{L0}\right)^2}{B \cdot f_L^2} \cdot v^{-1} + \frac{f_L - b_{L0}}{f_L^2} \tag{21}$$

D. Effective Stereo Baseline—Unfocused Versus Focused

In Section IV-A we showed that a focused plenoptic camera can be interpreted as an array of very narrow FOV virtual cameras with high resolution (full resolution or plenoptic 2.0 rendering). An unfocused plenoptic camera can be interpreted as an array of wide FOV virtual cameras with low resolution [35] (plenoptic 1.0 rendering). Here the resolution of a sub-aperture image, observing the complete scene, is limited by the number of micro lenses in the MLA.

Even though the images of both concepts have quite different characteristics, the camera setups differ only by the focal length

of the micro lenses. For some cases the plenoptic 1.0 and 2.0 concept can even be applied to the same raw image [33].

In the following we compare the effective stereo baseline for both concepts and therefore the benefits with respect to 3D scene reconstruction.

For plenoptic 1.0 rendering the maximum stereo baseline $\Delta B_{1.0\,\max}$ results from the distance B between MLA and sensor, the micro lens diameter D_M, and the main lens focal length f_L [35]:

$$\Delta B_{1.0\,\max} = \frac{D_M \cdot f_L}{B} \tag{22}$$

Since the virtual camera array is formed at distance f_L in front of the main lens, the object distance z_C can be calculated based on the pixel disparity μ in the sub-aperture images as given in eq. (23).

$$z_C = \frac{\Delta B_{1.0\,\max} \cdot f_L}{D_M \cdot \mu} + f_L \tag{23}$$

Here μ is scaled by the micro lens diameter D_M which defines the size of a pixel in the sub-aperture image. From eq. (23) we receive the depth accuracy $\sigma_{z_C\,1.0}$ for plenoptic 1.0 rendering given in eq. (24).

$$\sigma_{z_C\,1.0} = \left|\frac{\partial z_C}{\partial \mu}\right| \cdot \sigma_\mu = \frac{(z_C - f_L)^2}{f_L^2} \cdot B \cdot \sigma_\mu \tag{24}$$

Using the plenoptic camera interpretation of Section IV-A we can calculate the effective stereo baseline and therefore the theoretical depth accuracy for plenoptic 2.0 (full resolution) rendering in a similar way. The effective stereo baseline $\Delta B_{2.0}(\kappa)$ for a pair of projected micro images is received as follows:

$$\begin{aligned}\Delta B_{2.0}(\kappa) &= \|\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\| \\ &= \|\boldsymbol{c}_{ML}^{(k)} - \boldsymbol{c}_{ML}^{(l)}\| \frac{f_L}{b_{L0} - f_L} \\ &= \kappa \cdot D_M \cdot \frac{f_L}{b_{L0} - f_L}\end{aligned} \tag{25}$$

Here κ defines the multiple of D_M between the two micro lens centers. Therefore $\kappa \geq 1$ holds.

The object distance z_C is calculated based on the disparity μ_p in the projected micro images as given in eq. (26).

$$z_C = \frac{\Delta B_{2.0}(\kappa)}{\mu_p} - z_{C0} \tag{26}$$

Again, the depth accuracy $\sigma_{z_C\,2.0}$ is received from the standard deviation σ_μ of the pixel disparity as follows:

$$\begin{aligned}\sigma_{z_C\,2.0} &= \left|\frac{\partial z_C}{\partial \mu_p}\right| \cdot \sigma_{\mu_p} = \left|\frac{\partial z_C}{\partial \mu_p}\right| \cdot \left|\frac{\partial \mu_p}{\partial \mu}\right| \cdot \sigma_\mu \\ &= \frac{(z_C + z_{C0})^2 \cdot b_{L0}}{\kappa \cdot D_M \cdot z_{C0}} \cdot \frac{f_L - b_{L0}}{f_L \cdot B} \cdot s_{pixel} \cdot \sigma_\mu \\ &= \frac{(z_C + z_{C0})^2}{\kappa \cdot D_M} \cdot \frac{(f_L - b_{L0})^2}{f_L^2 \cdot B} \cdot s_{pixel} \cdot \sigma_\mu\end{aligned} \tag{27}$$

In eq. (27) $\left|\frac{\partial \mu_p}{\partial \mu}\right|$ defines the scaling from the pixel disparity μ to the disparity μ_p in the projected micro images. Here s_{pixel} defines the size of a pixel.

Using eq. (24) and eq. (27) we calculate the expected depth accuracies (standard deviations of the object distance z_C) for plenoptic 1.0 and 2.0 rendering. Fig. 4 shows the accuracies for both concepts, while using the same camera parameters ($f_L = 35$ mm, $B = 0.35$ mm, $b_{L0} = 34.3$ mm, $D_M = 0.1265$ mm, $s_{pixel} = 5.5$ μmm). For both rendering methods we chose $\sigma_\mu = 1$ pixel.

From eq. (24) and eq. (27) one can see that both curves are parabola shaped. Though, for the given setup $\sigma_{z_C\,1.0}$ has a much steeper slope than $\sigma_{z_C\,2.0}$ (see Fig. 4), which is due to a shorter effective stereo baseline.

For the shown setup ($b_{L0} < f_L$) the virtual camera array of the plenoptic 2.0 rendering lies behind the main lens, while the one for plenoptic 1.0 rendering lies always at distance f_L in front of the main lens. This leads to different minima in the curves and thus to an intersection of both curves, as can be seen from Fig. 4(b).

For smaller object distances the plenoptic 2.0 approach can use further apart micro lenses for stereo matching ($\kappa > 1$), which again leads to an improved accuracy.

Therefore, as can be seen from Fig. 4, the plenoptic 2.0 approach is always superior to the plenoptic 1.0 approach with respect to depth estimation. The κ-s shown in Fig. 4 are the first 10 received for a hexagonal arrangement of the MLA (1.00, 1.73, 2.00, 2.65, 3.00, 3.46, 3.61, 4.00, 4.36, 4.58).

We did not evaluate the functions for object distances closer than 0.5 m since here the images of both concepts can not be considered to be in focus anymore and thus the real accuracy will deviate from the curves. Anyway, in VO we are interested in larger object distances.

The only way to improve the depth accuracy of plenoptic 1.0 with respect to plenoptic 2.0 rendering is to reduce B. Though, this seems to be unfeasible due to the thickness of the MLA and the resulting impractical small F-number of the main lens (see F-number matching in [36]).

V. Plenoptic Camera Calibration

We present a plenoptic camera calibration approach which is based on the multiple view geometry introduced in Section IV-B. For calibration a complete bundle adjustment is performed, which optimizes the parameters of the plenoptic camera model, the 3D object coordinates and all camera poses at the same time. This is done based on recorded marker points, which are detected in the micro images of the camera.

Since we perform a complete bundle adjustment, we do not rely on any prior knowledge about the calibration target, like calibration points lying on a planar, regular checkerboard grid for instance. Furthermore, due to the 3D structure of our calibration target the optimization problem is better conditioned than those which are based on a planar target.

The calibration process determines the intrinsic parameters of the plenoptic camera, which are presented in Section IV-A, and additional distortion parameters. Our distortion model is

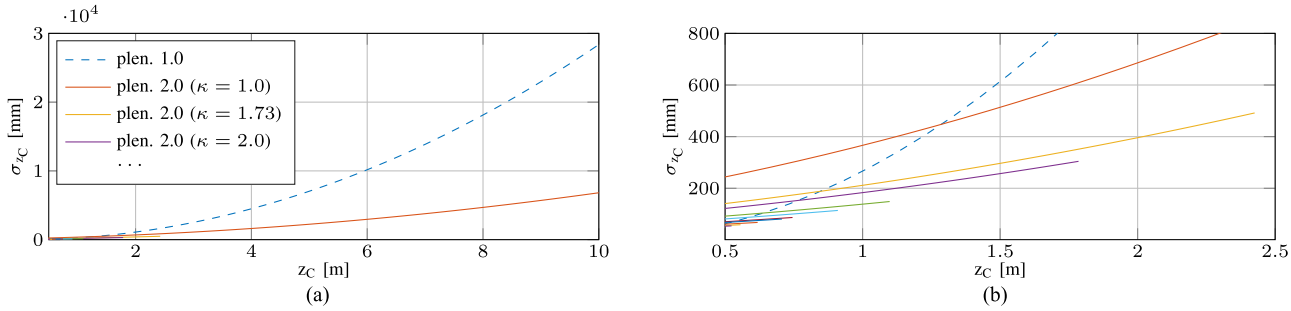


Fig. 4. Depth accuracy of plenoptic cameras based on plenoptic 1.0 and 2.0 rendering. Plenoptic 2.0 (full resolution) rendering results in a much larger stereo baseline and therefore better depth accuracy compared to plenoptic 1.0 approaches. (a) Depth accuracy. (b) Zoomed subsection of (a).

presented in Section V-A2. We do not consider the, in general marginal, effect of sensor tilting to preserve the planar arrangement of projected micro lens centers $\boldsymbol{p}_{ML}$ (as shown in Fig. 3). In Section V-B we present how to receive an initial solution for the bundle adjustment. The bundle adjustment, as we apply it to the calibration task, is presented in Section V-C.

A. Plenoptic Camera Model

1) Intrinsic Parameters: The plenoptic camera model consists of the following intrinsic parameters:

1) f_L – focal length of the main lens
2) b_{L0} – distance between MLA and main lens
3) B – distance between MLA and image sensor
4) $\boldsymbol{c}_L$ – principal point (intersection of the optical axis of the main lens with the image sensor)
5) $\boldsymbol{c}_{ML}^{(k)}$ – micro lens centers ($k \in \{1, 2, 3, \ldots\}$)

For the definition of these parameters we refer the reader to Section IV-B and Fig. 2.

To receive all intrinsic parameters in millimeters we define the pixel size, which we receive from the camera specification, as constraint. However, the correct pixel size is not relevant to obtain the optimum solution for the overall projection process.

While the micro lens centers $\boldsymbol{c}_{ML}^{(k)}$ are also part of the intrinsic camera parameters, we do not adjust them in the bundle adjustment. They can be easily estimated in advance using a white image (see [3] for instance). Furthermore, due to the small size of the micro images, in general we do not have more than one calibration point per micro image which would make the estimation of the micro lens centers in the bundle adjustment error prone.

2) Distortion Model: Most of the existing focused plenoptic camera models (i.e., [6], [7], [10]) define the projection from object space to virtual image space. Therefore, depth estimation has to be performed in advance using the distorted micro images and thus will be affected by the distortion of the main lens.

Since our model defines the projection from object space to the micro images, we are able to define the distortion model directly on the sensor before depth estimation. Hence, we do not have to deal with depth distortion afterwards.

Our distortion model considers radial and tangential distortion and is based on the well known model of Brown [37].

Here the radial distortion term Δr_{rad} is defined by a polynomial of the variable r, as given in eq. (28).

$$\Delta r_{rad} = A_0 r^3 + A_1 r^5 + A_2 r^7 + \cdots \tag{28}$$

The variable r is the distance between the principal point and the point coordinates on the MLA:

$$r = \sqrt{x^2 + y^2} \tag{29}$$

This results in the Cartesian correction terms Δx_{rad} and Δy_{rad} as given in eq. (30) and (31).

$$\Delta x_{rad} = x \frac{\Delta r_{rad}}{r} = x \left(A_0 r^2 + A_1 r^4 + A_2 r^6 + \cdots \right) \tag{30}$$

$$\Delta y_{rad} = y \frac{\Delta r_{rad}}{r} = y \left(A_0 r^2 + A_1 r^4 + A_2 r^6 + \cdots \right) \tag{31}$$

The tangential distortion terms (Δx_{tan} and Δy_{tan}) are defined as given in eq. (32) and (33).

$$\Delta x_{tan} = B_0 \left(r^2 + 2x^2 \right) + 2 B_1 xy \tag{32}$$

$$\Delta y_{tan} = B_1 \left(r^2 + 2y^2 \right) + 2 B_0 xy \tag{33}$$

Based on the correction terms the distorted coordinates $\tilde{x}$ and $\tilde{y}$ are calculated from the ideal projection as follows:

$$\tilde{x} = x + \Delta x_{rad} + \Delta x_{tan} \tag{34}$$

$$\tilde{y} = y + \Delta y_{rad} + \Delta y_{tan} \tag{35}$$

This model has the nice property that it consists only of additive distortion terms which depend on the undistorted coordinates. Due to the small size of a micro image it is sufficient to consider the distortion to be constant within a single micro image and therefore we only have to correct the micro image centers $\boldsymbol{c}_{ML}$ and not the point coordinates $\boldsymbol{x}_{ML}$ in the respective micro image.

By construction the micro lenses form a regular grid. The coordinates of their centers, which are estimated cf. [3] are the distorted micro lens centers $\tilde{\boldsymbol{c}}_{ML}$. Therefore, the micro lens centers $\boldsymbol{c}_{ML}$ which are corrected during the calibration considering the lens distortion and which are used for the projection, will deviate from this regular grid.

B. Initialization of the Calibration Process

While one is able to build a closed form solution for a SfM problem based on an uncalibrated monocular camera, this is not

the case for a plenoptic camera. The reason is that for each pair of micro images from distinct views one obtains a fundamental matrix. To estimate this one would need multiple corresponding points in each micro image, which, considering the small size of the micro images, is not realistic. This makes the initialization of the optimization problem more difficult. To initialize the bundle adjustment we consider in a first step that the totally focused image of the plenoptic camera results from a regular pinhole camera. In this way we solve the SfM problem using a standard photogrammetric software. For initialization this approximation is sufficient since only a rough estimate of all parameters is needed.

In addition, initial parameters for B and b_{L0} are received by solving the physical model defined in [9]. In this way we receive initial values for all intrinsic and extrinsic camera parameters as well as the 3D object points. A detailed description of the initialization can be found in [10].

Beside the initial parameters, the photogrammetric software provides already the correct correspondences of the recorded calibration points.

In the following we proof the validity of using the pinhole camera model for initialization, at least for $f_L \ll z_C$, based on a specific example. The camera setup is as follows:

1) main lens focal length $f_L = 16$ mm
2) size of the totally focused image: 1024×1024 pixel
3) principal point in the image center

The calibration target covers a depth range from $z_{C\,\min} = 500$ mm up to $z_{C\,\max} = 1500$ mm. Based on the thin lens equation we are able to calculate the following maximum, minimum and average image distances:

1) $b_{L\,\max} = 16.53$ mm
2) $b_{L\,\min} = 16.17$ mm
3) $\bar{b}_L = 16.35$ mm

Thus, one receives a maximum projection error between the plenoptic and pinhole camera model as follows:

$$\begin{aligned}\Delta x_{\max} &= x_{\max} \cdot \left(\frac{b_{L\,\max}}{\bar{b}_L} - 1 \right) \\ &= x_{\max} \cdot \left(1 - \frac{b_{L\,\min}}{\bar{b}_L} \right) = 5.63 \text{ pixel} \end{aligned} \tag{36}$$

Here the image plane of the pinhole camera model is considered to be in distance $\bar{b}_L$ to the main lens. The maximum image coordinate $x_{\max}$ is considered to be at the image boundary and therefore at $x_{\max} = 512$ pixel. As can be seen from eq. (36) the error made for the initialization is only in the range of a few pixels.

For larger object distances the error further decreases (i.e., for $z_{C\,\min} = 1500$ mm and $z_{C\,\max} = 2500$ mm it follows $\Delta x_{\max} = 1.11$ pixel). The variation in the average image distance $\bar{b}_L$ results in a small scaling error of the recorded image and therefore a bias in the estimated object distance. For $f_L \ll z_C$ it follows $b_L \approx f_L$ and therefore the estimated principal distance of the pinhole camera model gives a good initialization for the main lens focal length f_L of the plenoptic camera.

C. Plenoptic Bundle Adjustment

To estimate the exact camera parameters we formulate the following non-linear optimization problem:

$$E(\Pi, \Xi, P) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{L} \left\| \boldsymbol{r}_{(i,j,k)} \right\|^2 \cdot \theta_{(i,j,k)} \tag{37}$$

Here, Π is the set of all camera parameters (intrinsic and distortion parameters), Ξ is the set of all camera poses ($\Xi := \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M\}$) and P the set of all object point coordinates ($P := \{\boldsymbol{x}_W^{(1)}, \ldots, \boldsymbol{x}_W^{(N)}\}$). Each calibration point in a micro image results in a 2D residual vector $\boldsymbol{r}_{(i,j,k)}$. $\theta_{(i,j,k)}$ is a masking function which is 1 if the i-th object point is visible in the k-th micro image of the j-th view and zero otherwise.

The residual vector for the calibration point with coordinates $\boldsymbol{x}_{ML}$, corresponding to the i-th object point which is seen in the k-th micro image of the j-th camera view, is defined as follows:

$$\boldsymbol{r}_{(i,j,k)} = \pi_{ML} \left(\boldsymbol{G}(\boldsymbol{\xi}_j) \boldsymbol{x}_W^{(i)}, \boldsymbol{c}_{ML}^{(k)}, \Pi \right) - \boldsymbol{x}_{ML} \tag{38}$$

The function $\pi_{ML}(\cdot)$ defines the projection from camera to micro image coordinates of the k-th micro image. In eq. (37) the sum over i is the sum over all object points, the sum over j is the sum over all camera poses and the sum over k is the sum over all micro images. To become robust against outliers, $\theta_{(i,j,k)}$ can also be extended with any robust loss function. The vector $\boldsymbol{\xi}_j \in \mathfrak{se}(3)$ defines the rigid body transformation $\boldsymbol{G}(\boldsymbol{\xi}_j) \in \mathrm{SE}(3)$ from world coordinates to camera coordinates of the j-th view.

The optimal parameters ($\hat{\Pi}, \hat{\Xi}, \hat{P}$) are the one which minimize the cost function $E(\Pi, \Xi, P)$:

$$\{\hat{\Pi}, \hat{\Xi}, \hat{P}\} = \arg\min E(\Pi, \Xi, P) \tag{39}$$

We solve this highly nonlinear optimization problem using the Levenberg-Marquardt algorithm. The initialization of Π, Ξ and P has to guarantee that the optimization starts in the convex region around the optimum solution of $E(\Pi, \Xi, P)$.

The scaling of the object is received based on known distances between certain object points. Those distances are used as additional constraints in the optimization problem.

Our implementation of the plenoptic camera calibration is based on the Ceres Solver Library [38].

VI. Direct Plenoptic Odometry

Our DPO algorithm performs direct image alignment and semi-dense mapping directly on the micro images recorded by a plenoptic camera. It makes use of the geometric camera interpretation defined in Section IV. The geometric camera interpretation relies on known intrinsic camera parameters, as well as distortion corrected images. Therefore, in advance to run DPO, the plenoptic camera has to be calibrated as described in Section V.

A. Overview

Algorithm 1 shows the complete workflow. We give here a short overview and focus on the single steps in the following subsections (Sections VI-B–VI-F). A new recorded light-field frame is tracked based on a keyframe. Keyframes

Algorithm 1: Direct Plenoptic Odometry.

Initialization:
currenKeyFrame $\leftarrow$ get first light-field frame in sequence $I_{ML0}(\boldsymbol{x}_R)$;
/* frame initially consists only of raw intensity image $I_{ML}(\boldsymbol{x}_R)$ */
$D_{ML}(\boldsymbol{x}_R) \leftarrow$ do In-Frame depth estimation on *currentKeyFrame*; /* see Sec. VI-C2 */
$D_V(\boldsymbol{x}_V) \leftarrow$ calculate central perspective depth map from $D_{ML}(\boldsymbol{x}_R)$;
$D_V(\boldsymbol{x}_V) \leftarrow$ regularize depth map $D_V(\boldsymbol{x}_V)$ and remove outliers; /* see Sec. VI-C4 */
$D_{ML}(\boldsymbol{x}_R) \leftarrow$ remove outliers in $D_{ML}(\boldsymbol{x}_R)$ which were detected in $D_V(\boldsymbol{x}_V)$;
$I_V(\boldsymbol{x}_V) \leftarrow$ calculate totally focused image for *currentKeyFrame* using $D_V(\boldsymbol{x}_V)$ and $I_{ML}(\boldsymbol{x}_R)$;
/* indices of current keyframe depth maps and intensity images are ignored since they exist only once */
for $j = 1 \rightarrow$ end of sequence **do**
  currentFrame $\leftarrow$ get next light-field frame in sequence $I_{MLj}(\boldsymbol{x}_R)$;
  Tracking:
  $\{\boldsymbol{\xi}_{kj},\ pointUsage,\ trackingResidual\} \leftarrow$ estimate pose from *currentKeyFrame* to *currentFrame* using $D_V(\boldsymbol{x}_V)$, $I_V(\boldsymbol{x}_V)$, $I_{MLj}(\boldsymbol{x}_R)$ and $\boldsymbol{\xi}_{k(j-1)}$; /* see Sec. VI-E */
  Depth Estimation:
  $D_{ML}(\boldsymbol{x}_R) \leftarrow$ estimate Inter-Frame depth from *currentKeyFrame* to *currentFrame* using $I_{ML}(\boldsymbol{x}_R)$, $D_{ML}(\boldsymbol{x}_R)$, $I_{MLj}(\boldsymbol{x}_R)$ and $\boldsymbol{\xi}_{kj}$; /* see Sec. VI-C3 */
  /* new depth estimates are merged into existing depth map $D_{ML}(\boldsymbol{x}_R)$ as described in Sec. VI-D */
  $D_V(\boldsymbol{x}_V) \leftarrow$ recalculate central perspective depth map from $D_{ML}(\boldsymbol{x}_R)$;
  $D_V(\boldsymbol{x}_V) \leftarrow$ regularize depth map $D_V(\boldsymbol{x}_V)$ and remove outliers; /* see Sec. VI-C4 */
  $D_{ML}(\boldsymbol{x}_R) \leftarrow$ remove outliers in $D_{ML}(\boldsymbol{x}_R)$ which were detected in $D_V(\boldsymbol{x}_V)$;
  $I_V(\boldsymbol{x}_V) \leftarrow$ recalculate totally focused image for *currentKeyFrame* using $D_V(\boldsymbol{x}_V)$ and $I_{ML}(\boldsymbol{x}_R)$;
  Keyframe Selection:
  newKeyframeScore $\leftarrow$ calculate new keyframe score using $\boldsymbol{\xi}_{kj}$ and *pointUsage*; /* see Sec. VI-F */
  if *newKeyframeScore* $>$ *keyframeSelectionThreshold* **then**
    newKeyFrame $\leftarrow$ *currentFrame*;
    $D_{ML\text{tmp}}(\boldsymbol{x}_R) \leftarrow$ estimate In-Frame depth on *newKeyFrame*; /* see Sec. VI-C2 */
    $D_{ML}(\boldsymbol{x}_R) \leftarrow$ propagate $D_{ML}(\boldsymbol{x}_R)$ from *currentKeyFrame* to *newKeyFrame* using $\boldsymbol{\xi}_{kj}$;
    $D_{ML}(\boldsymbol{x}_R) \leftarrow$ merge $D_{ML}(\boldsymbol{x}_R)$ with $D_{ML\text{tmp}}(\boldsymbol{x}_R)$;
    currentKeyFrame $\leftarrow$ *newKeyFrame*;
    $D_V(\boldsymbol{x}_V) \leftarrow$ recalculate central perspective depth map from $D_{ML}(\boldsymbol{x}_R)$;
    $D_V(\boldsymbol{x}_V) \leftarrow$ regularize depth map $D_V(\boldsymbol{x}_V)$ and remove outliers; /* see Sec. VI-C4 */
    $D_{ML}(\boldsymbol{x}_R) \leftarrow$ remove outliers in $D_{ML}(\boldsymbol{x}_R)$ which were detected in $D_V(\boldsymbol{x}_V)$;
    $I_V(\boldsymbol{x}_V) \leftarrow$ recalculate totally focused image for *currentKeyFrame* using $D_V(\boldsymbol{x}_V)$ and $I_{ML}(\boldsymbol{x}_R)$;
  end
end

are selected light-field frames in respect to which tracking and mapping of all recorded frames takes place. In addition to the light-field (raw image), for each keyframe two depth maps (micro image depth map and virtual image depth map) are established and a totally focused intensity image is synthesized. After successful tracking of a light-field frame the current keyframe is updated, i.e., the depth maps are refined based on the tracked frame and the totally focused image is recalculated.

During progress of the DPO algorithm, the pose of newly recorded frames differs more and more from the one of the initially established keyframe. Based on a score we check for each new frame whether the old keyframe is still valid or whether the new frame has to be set as a new keyframe. For a new keyframe, in-frame depth estimation is performed based on its recorded light-field. The in-frame depth map is merged with the one propagated from the last keyframe. Hence only the depth maps of the current keyframe will be kept.

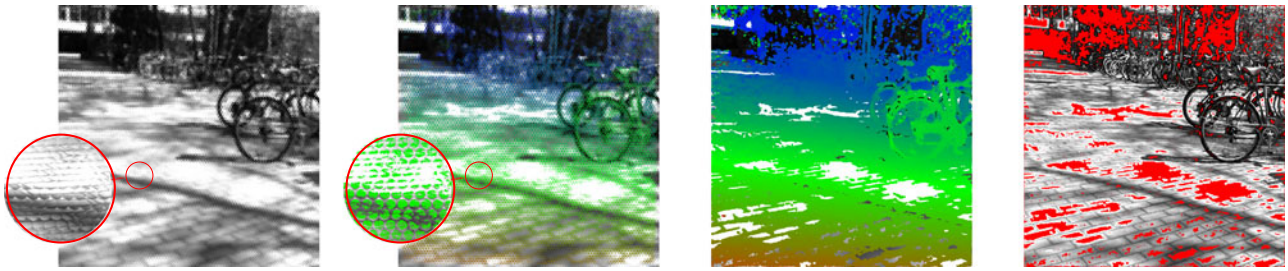
B. Depth Map Representation

For a plenoptic camera, the observed inverse virtual depth v^{-1} can be considers to result from a Gaussian process [34]. Therefore, due to eq. (21) the observed inverse effective depth $\lambda^{-1} = z_C'^{-1}$ has to result form a Gaussian process, too. Similar to [18] we define for each image point with an absolute gradient above a certain threshold a probabilistic depth hypothesis. We model the inverse effective depth as a Gaussian process:

$$\mathcal{N}\left(d, \sigma_d^2\right) \tag{40}$$

Here, $d = z_C'^{-1}$ defines the mean, while σ_d^2 is the corresponding variance.

1) Micro image Depth Map and Virtual Image Depth Map: Using the probabilistic depth model (eq. (40)) two different depth maps ($D_{ML}(\boldsymbol{x}_R)$ and $D_V(\boldsymbol{x}_V)$) are defined for each keyframe (Fig. 5(b) and (c)).



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

Fig. 5. Keyframe depth maps and intensity images. (a) Micro images $I_{ML}(\boldsymbol{x}_R)$ (raw image recorded by the plenoptic camera). (b) Micro image depth map $D_{ML}(\boldsymbol{x}_R)$. (c) Virtual image depth map $D_V(\boldsymbol{x}_V)$. (d) Totally focused intensity image $I_V(\boldsymbol{x}_V)$. For the pixels marked in red no depth value and hence no intensity was calculated.

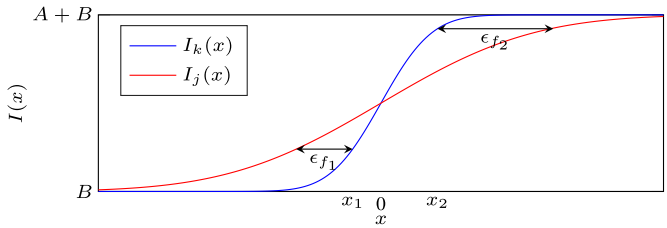


Fig. 6. Focus disparity error ϵ_f for two different positions (x_1 and x_2) on an edge along the epipolar line.

$D_{ML}(\boldsymbol{x}_R)$ is defined on raw image coordinates $\boldsymbol{x}_R$. Here a point in the i-th micro image, with micro image coordinates $\boldsymbol{x}_{ML}$ has the raw image coordinates $\boldsymbol{x}_R = \boldsymbol{x}_{ML} + \boldsymbol{c}_{ML}^{(i)}$. $D_V(\boldsymbol{x}_V)$ as well as a totally focused intensity image $I_V(\boldsymbol{x}_V)$ (Fig. 5(d)) are defined on virtual image coordinates $\boldsymbol{x}_V$ [34], where micro image points observing the same object point are merged (see Section VI-D).

We consider the two distinct depth maps because in the raw image one object point is projected to multiple image points, while there is a one to one mapping from object space to the virtual image.

Depth estimation is performed on $D_{ML}(\boldsymbol{x}_R)$, while the depth map $D_V(\boldsymbol{x}_V)$ is used for tracking new frames.

C. Estimating Depth Hypotheses

For each keyframe we perform in-frame depth estimation based on stereo matching in its own micro images as well as inter-frame depth estimation based on micro images of subsequent frames using the epipolar gemometry defined in Section IV-B. For both in-frame and inter-frame depth estimation stereo matches are found by optimizing the sum of squared intensity difference (SSID) over a one-dimensional pixel patch along the epipolar line. Prior estimates are used to narrow the search range to $d \pm 2\sigma_d$.

1) Observation Uncertainty: Similar to [18] we derive the observation uncertainty, which results in the variance σ_d^2 of the inverse effective depth d, based on uncertainty propagation. In addition to a geometric and a photometric disparity error, a focus disparity error, which regards unfocused micro images, is defined.

a) Geometric disparity error: Due to inaccuracies in the camera model and uncertainties in the frame pose one can expect a misalignment of the epipolar line. Due to the very short search range, we consider, similar to [18], only the absolute epipolar line position $\boldsymbol{l}_0$ to be effected by isotropic Gaussian noise ϵ_l and we can neglect any rotational error. Since for plenoptic cameras the search range is always limited by the micro lens dimension, the assumption of a short search range holds even better than for the monocular case. Thus, the variance $\sigma^2_{\mu(\xi,\pi)}$ of the geometric disparity error can be defined as follows [18]:

$$\sigma^2_{\mu(\xi,\pi)} = \frac{\sigma_l^2}{\langle \boldsymbol{g}, \boldsymbol{l} \rangle^2} \tag{41}$$

Here, $\boldsymbol{g}$ is the normalized image gradient, $\boldsymbol{l}$ the normalized epipolar line direction, and σ_l^2 the variance of ϵ_l.

b) Photometric disparity error: The photometric disparity error describes the effect of sensor noise ϵ_n on the estimated disparity. It results in an error on the estimated disparity with variance $\sigma^2_{\mu(I)}$, defined as follows [18]:

$$\sigma^2_{\mu(I)} = \frac{2 \cdot \sigma_n^2}{g_I^2} \tag{42}$$

Here σ_n^2 is the variance of ϵ_n and g_I the intensity gradient along the epipolar line.

c) Focus disparity error: Different to regular cameras, which in general are focused to infinity, the micro images of the plenoptic camera can not be considered to be in focus for the complete operating range, especially not for a multi-focus plenoptic camera [36]. Hence, the focusing itself also affects the stereo observation (Fig. 6). In the following we derive the variance $\sigma^2_{\mu(v,k,j)}$ of the focus disparity error.

Let k be the index of the micro image for which mapping is performed, while j is the one of the stereo reference. To model the focus disparity error we consider the real edge $I_{real}(x)$ which is observed in the micro images as a perfect Heaviside-step-function $h(x)$ with amplitude A and offset B along the epipolar line:

$$I_{real}(x) = A \cdot h(x) + B \tag{43}$$

The variable x is the position on the respective epipolar line relative to the step position μ_{0i} ($i \in \{k, j\}$):

$$x = \mu_k - \mu_{0k} = \mu_j \cdot \gamma - \mu_{0j} \tag{44}$$

Therefore, the correct disparity is defined by $\mu^* = \mu_{0j} - \mu_{0k}$. The parameters $\gamma = \frac{z_C'^{(j)}}{z_C'^{(k)}}$ defines the scaling factor between the micro images. During the imaging process the edge $I_{real}(x)$ is filtered by a Gaussian filter with variance σ_i^2 which depends on the virtual depth v, the micro lens type and the sampling rate (pixel pitch).

Thus, on the sensor we receive the following intensity functions along the epipolar line:

$$I_i(x) = B + \frac{A}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sigma_i\sqrt{2}}\right)\right) \quad i \in \{k, j\} \tag{45}$$

The estimated disparity $\hat{\mu} = \mu^* + \epsilon_f$ is the one for which both intensities have the same value and therefore, the following condition is fulfilled:

$$I_k(x) \stackrel{!}{=} I_j(x - \epsilon_f) \tag{46}$$

$$\operatorname{erf}\left(\frac{x}{\sigma_k\sqrt{2}}\right) \stackrel{!}{=} \operatorname{erf}\left(\frac{x - \epsilon_f}{\sigma_j\sqrt{2}}\right) \tag{47}$$

Since the error function ($\operatorname{erf}(\cdot)$) can not be solved analytically, we linearize eq. (47) and receive, after a rearrangement, the following relationship between the focus disparity error ϵ_f and the position on the edge x:

$$\epsilon_f = x \cdot \left(1 - \frac{\sigma_j}{\sigma_k}\right) \tag{48}$$

Fig. 6 visualizes the focus disparity error ϵ_f exemplary for two different positions (x_1 and x_2) on an edge along the epipolar line. Here, the red and blue curves represent the intensity along the epipolar line in two different micro images with different blur radii. For a certain position x on the edge the estimated disparity $\hat{\mu}$ will be the one for which both curves have the same value and therefore will be shifted by ϵ_f with respect to the real disparity μ^* as given in eq. (46). For the case that both images have the same blur radius, both curves are perfectly overlaid and therefore $\epsilon_f = 0$ will hold for any position x.

Considering the position with respect to the real edge x as a random variable with variance σ_x^2, the variance $\sigma_{\mu(v,k,j)}^2$ of the focus disparity error is as follows:

$$\sigma_{\mu(v,k,j)}^2 = \sigma_x^2 \cdot \left(1 - \frac{\sigma_j}{\sigma_k}\right)^2 \tag{49}$$

The standard deviation σ_i ($i \in \{k, j\}$) depends on the blur diameter s_i of the respective micro image, which is calculated based on the virtual depth v and the micro lens parameters (D_M, f_M, B) using the thin lens equation:

$$s_i = D_M \cdot \left|\frac{1}{v_i} + \frac{B}{f_{Mi}} - 1\right| \tag{50}$$

Since the minimum blur radius is limited by the pixel pitch and the overall optical system, the blur radius has a lower boundary s_0. Thus the following variances σ_k^2 and σ_j^2 result:

$$\sigma_k^2 = \beta^2 \cdot \min\{s_k^2, s_0^2\}, \quad \sigma_j^2 = \beta^2\gamma^2 \cdot \min\{s_j^2, s_0^2\} \tag{51}$$

The constant parameter β models the scaling from blur diameter to the standard deviation of the Gaussian filter.

Considering all three error sources as independent random variables, the overall observation uncertainty is received as follows:

$$\sigma_d^2 = \alpha^2 \cdot \left(\sigma_{\mu(\xi,\pi)}^2 + \sigma_{\mu(I)}^2 + \sigma_{\mu(v,k,j)}^2\right) \tag{52}$$

where α defines the derivative of d with respect to the disparity μ.

2) Estimating In-Frame Depth

In-frame depth estimation is performed similarly to the procedure we have described previously in [34]. The only difference is that here we directly estimate the inverse effective depth $z_C'^{-1}$ instead of the inverse virtual depth v^{-1}. Nevertheless, since there is a linear connection, the overall procedure stays the same.

3) Estimating Inter-frame Depth

While for in-frame depth estimation the stereo baseline is limited by the camera dimensions, we are able to improve the depth accuracy based on inter-frame depth observations. Inter-frame depth estimation is performed from the current keyframe to newly tracked frames. Therefore, we consider the relative orientation between the frames $\boldsymbol{\xi} \in \mathfrak{se}(3)$ to be known.

For each micro image point in the keyframe stereo observations are obtained from all possible micro images in the new frame.

a) Defining epipolar lines: One is able to define the epipolar geometry between the micro images of two different frames as given in Section IV-B. Due to the linear relation between projected micro image coordinates $\boldsymbol{x}_p$ and real micro image coordinates on the sensor $\boldsymbol{x}_{ML}$ (or $\boldsymbol{x}_R$), the epipolar lines defined in the projected micro images can be directly mapped on the sensor. Therefore, stereo matching for inter-frame depth estimation can be performed directly on the recorded raw images.

4) Regularizing Depth Maps

Each time the depth maps are updated a regularization step on $D_{ML}(\boldsymbol{x}_R)$ and $D_V(\boldsymbol{x}_V)$ is performed. Here, outliers are removed and estimates are smoothed based on probabilistic metrics, similar to [39]. Since it is much easier to detect outliers in $D_V(\boldsymbol{x}_V)$ the corresponding points in $D_{ML}(\boldsymbol{x}_R)$ are marked as outliers, too. The depth hypotheses in $D_{ML}(\boldsymbol{x}_R)$ itself are not updated based on $D_V(\boldsymbol{x}_V)$ to avoid loops in the filtering process. After updating $D_V(\boldsymbol{x}_R)$, the totally focused intensity image $I_V(\boldsymbol{x}_R)$ is recalculated, too.

D. Merging Depth Hypotheses

Probabilistic depth values are updated in a Kalman-like fashion, similar to [18]. However, in contrast to [18] we limit the variance of the merged depth hypothesis to the variance of the best observation. By this we do not consider the observations to be uncorrelated, which in the limit case would lead to zero variance. Thus, the following merging routine is defined:

$$\mathcal{N}\left(\frac{\sum_{i\in O} d_i \cdot \left(\sigma_{di}^2\right)^{-1}}{\sum_{i\in O}\left(\sigma_{di}^2\right)^{-1}}, \min\left(\sigma_{di}^2\right)\right) \tag{53}$$

where O is the set of depth observations. This algorithm avoids that points in the background which are observed in numerous frames receive low variances.

E. Tracking

Newly recorded frames are tracked based on direct image alignment. As tracking reference we use the current keyframe ($D_V(\boldsymbol{x}_V)$ and $I_V(\boldsymbol{x}_V)$).

Tracking is performed in two steps. Initial tracking is performed based on a coarse-to-fine approach using the totally focused image of the new frame. Final tracking is performed directly on the micro images.

1) Initial Tracking on Pyramid Levels

To obtain the totally focused image of the new frame without performing in-frame depth estimation, the virtual image depth map of the current keyframe is propagated to the new frame based on the pose of the last tracked frame. With this depth information the totally focused image can be synthesized.

For initial tracking we perform direct image alignment based on a coarse-to-fine approach as proposed in [40] to increases the radius of convergence. With the Levenberg-Marquardt algorithm we optimize the variance-normalized intensity error $E_V(\boldsymbol{\xi}_{kj})$:

$$E_V(\boldsymbol{\xi}_{kj}) = \sum_i \left\| \left(\frac{r_V^{(i)}}{\sigma_r^{(i)}} \right)^2 \right\|_\delta \tag{54}$$

$$r_V^{(i)} := I_{Vk}\left(\boldsymbol{x}_V^{(i)}\right) - I_{Vj}\left(\pi_V\left(\boldsymbol{G}(\boldsymbol{\xi}_{kj})\pi_V^{-1}(\boldsymbol{x}_V^{(i)})\right)\right) \tag{55}$$

Here $\pi_V(\cdot)$ defines the projection from camera to virtual image coordinates, while $\pi_V^{-1}(\cdot)$ is the inverse projection. $\|\cdot\|_\delta$ is the robust Huber norm [41].

2) Tracking on Micro Images

The result from the initial tracking is used to initialize the micro image based tracking. Here, each point in the tracking reference (frame k) $\boldsymbol{x}_C^{(i)}$ is projected to all micro images in the new frame (frame j) which observe this point. This results in the following optimization function which is solved with respect to $\boldsymbol{\xi}_{kj}$:

$$E(\boldsymbol{\xi}_{kj}) = \sum_i \sum_l \left\| \left(\frac{r_{ML}^{(i,l)}}{\sigma_r^{(i,l)}} \right)^2 \right\|_\delta \tag{56}$$

$$r_{ML}^{(i,l)} := I_{Vk}\left(\boldsymbol{x}_V^{(i)}\right) - I_{MLj}\left(\pi_{ML}\left(\boldsymbol{G}(\boldsymbol{\xi}_{kj})\pi_V^{-1}(\boldsymbol{x}_V^{(i)}), \boldsymbol{c}_{ML}^{(l)}\right)\right) \tag{57}$$

Here $\pi_{ML}(\cdot)$ defines the projection from camera to micro image coordinates. By the projection of the reference points to multiple micro images in the new frame we are able to implicitly incorporate the in-frame disparities of the new frame into the optimization problem. In this way we are able to avoid accumulated scale drifts as they in general occur for monocular tracking.

3) Variance of Tracking Residual

For both initial and final tracking the residual variance σ_r^2 is received based on uncertainty propagation. Therefore, σ_r^2 consists of a photometric component, which results from noise on the intensities and a geometric component which results from noise on the depth estimate.

$$\sigma_r^2 := \sigma_n^2 \left(\frac{1}{N_k} + \frac{1}{N_j} \right) + \left| \frac{\partial r(\boldsymbol{x}_V, \boldsymbol{\xi}_{kj})}{\partial d(\boldsymbol{x}_V)} \right|^2 \sigma_d^2(\boldsymbol{x}_V) \tag{58}$$

We consider the noise on the intensities in the different micro images to be uncorrelated. Therefore, the variance of an intensity value is estimated by the variance of the sensor noise σ_n^2 divided by the number of micro image points used to calculate the intensity value:

$$\left(\sigma_I^{(i)}\right)^2 = \frac{\sigma_n^2}{N_i} \quad \text{with} \quad i \in \{k, j\} \tag{59}$$

Here N_i ($i \in \{k, j\}$) is the number of micro image points used for calculation. Hence, for the tracking on the micro images it follows $N_j = 1$.

The geometric term is received based on uncertainty propagation from the variance σ_d^2 of the corresponding depth hypothesis.

F. Selecting Keyframes

A new keyframe has to be selected when the view of a new frame differs too much from the one of the current keyframe. Therefore, we define a score based on the translation between the two frames as well as on the percentage of good points during tracking.

When a frame is selected to be the new keyframe, in-frame depth estimation is performed in the new frame. Afterwards, the micro image depth map of the last keyframe is projected into the new one and the depth hypotheses are merged. Here, projected depths which are divergent from the in-frame depth are rejected and not merged.

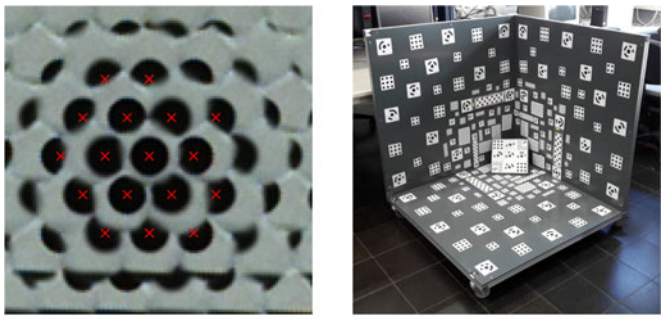
VII. Evaluation

A. Plenoptic Camera Calibration

In this section the evaluation of the calibration approach for the plenoptic camera is presented. We compare the calibration results for different main lenses. Besides, we compare our approach to a state-of-the-art calibration algorithm [5] for MLA based light-field cameras.

All experiments were performed with a Raytrix R5 camera (sensor size: 11,264 × 11,264 mm, image size: 2048 × 2048 pixel, diameter of micro lenses: ∼23 pixel, aperture: f/2.4).

Due to the reason that it is quite difficult to detect marker points directly in the micro images of the plenoptic camera, we detect them in our approach in the totally focused image and project them back to the micro images. Anyway, the totally focused image has to be calculated in advance to initialize the



(a) (b)

Fig. 7. 3D target used for plenoptic camera calibration. (a) Single marker point projected to multiple micro images. Detected marker coordinates are visualized by red crosses. (b) Complete 3D calibration target. Consists of (unique) coded and uncoded markers. Six defined distances between pairs of coded markers are used to retrieve the scale of the calibration object. (a) Marker point. (b) Calibration target.

TABLE II
ESTIMATED INTRINSIC CAMERA PARAMETERS AND REPROJECTION ERRORS FOR THREE DIFFERENT MAIN LENS FOCAL LENGTHS

Lens	f_L [mm]	b_{L0} [mm]	B [mm]	c_{Lx} [pixel]	c_{Ly} [pixel]	s_x [pixel]	s_y [pixel]
12.5 mm	12.616	11.789	0.353	1003.3	1043.9	0.213	0.220
16mm	16.273	15.482	0.357	1015.7	1056.3	0.123	0.126
35mm	34.868	33.993	0.367	1024.0	1042.4	0.063	0.067

bundle adjustment. Fig. 7(a) shows an example for the detected marker points in the micro images, while Fig. 7(b) shows the complete calibration target. In contrast, the method presented by Bok *et al.* [5] detects line features from a checkerboard pattern directly in the micro images.

1) Calibration With Different Main Lenses: We use three different main lenses with different focal lengths ($f_L = 12,5$ mm, $f_L = 16$ mm, $f_L = 35$ mm) to evaluate our calibration approach.

For each lens we estimate the camera parameters and we evaluate the reprojection error as well as the accuracy of the 3D points estimated during the bundle adjustment. This allows us to assess the validity of our model and the robustness of the calibration approach.

Table II shows the estimated intrinsic parameters for all three lenses and the corresponding reprojection errors. Here, s_x and s_y are the root mean square (RMS) values of the reprojection error in the x- and y-coordinate respectively.

From Table II one can see that for all three main lenses the reprojection errors are much smaller than one pixel. This confirms the validity of the complete projection model. Here it seems that the reprojection error is correlated with the main lens focal length f_L. Another indication for the validity of the defined projection model is that the estimated main lens focal length f_L is quite close to the nominal focal length of the respective lens. In addition, the estimated parameter B is similar for all three lenses. The parameter B is a constant of the plenoptic sensor and therefore does not depend on the main lens.

TABLE III
ACCURACY OF 3-D OBJECT COORDINATES ESTIMATED DURING THE BUNDLE ADJUSTMENT

Lens	12,5 mm	16 mm	35 mm
RMSE	0.289 mm	0.301 mm	0.332 mm
MAE	2.124 mm	2.352 mm	2.411 mm

TABLE IV
ROBUSTNESS OF THE ESTIMATED INTRINSIC CAMERA PARAMETERS USING A MAIN LENS FOCAL LENGTH OF $f_L = 16$ MM

Calibration Method		Our Method mean	st. dev.	rel. std.	Bok *et al.* [5] mean	std. dev.	rel. std [%]
f_L	[mm]	16.28	0.003	0.020%	–	–	–
b_{L0}	[mm]	15.53	0.018	0.118%	–	–	–
B	[mm]	0.36	0.007	1.975%	–	–	–
c_{Lx}	[pixel]	1015.07	0.705	0.069%	1000.09	13.651	1.365%
c_{Ly}	[pixel]	1046.89	0.390	0.037%	1048.73	11.051	1.054%
f_x	[pixel]	*2888.97*	*2.072*	*0.072%*	2963.39	6.406	0.216%
f_y	[pixel]	*2888.97*	*2.072*	*0.072%*	2963.01	6.577	0.222%
K_1		*2.04*	*0.007*	*0.319%*	1.93	0.094	4.907%
K_2		*689.49*	*14.917*	*2.164%*	767.41	20.988	2.735%

To evaluate the accuracy of the estimated 3D object coordinates, we register the point clouds received from the bundle adjustment for all three main lenses with a reference point cloud using ICP (iterative closest point). The reference point cloud is received based on SfM using a standard monocular camera. For this we used a professional photogrammetric measurement software to estimate a highly accurate reference point cloud. Table III shows the root mean square error (RMSE) and the maximum absolute error (MAE) between the point clouds received from the plenoptic camera based bundle adjustment and the reference point cloud. For all three main lenses we measure a RMSE of the point cloud of less than 1mm and thereby confirm the validity of the defined multiple view geometry. The MAE is less than 2.5 mm for all three lenses although also the outliers are considered.

A point cloud accuracy of less than 1mm is way less than the measurement accuracy of the plenoptic camera and therefore sufficient for a precise calibration of the camera.

2) Comparison to State of the Art: We compare our calibration approach to the one presented by Bok *et al.* [5]. Both methods define the complete projection from object space to the recorded micro images on the sensor. While the overall projection from object space to the micro images is quite similar for both models we mainly want to demonstrate that the robustness of the calibration significantly benefits from the 3D calibration target in comparison to the planar checkerboard used in [5].

To compare the methods, we recorded for both targets 64 images from different perspectives using the main lens with $f_L = 16$ mm.

To evaluate the robustness of the calibration approach we performed the calibration for each method 10 times by picking randomly a set of 20 images out of the complete collection of

TABLE V
TRACKING DRIFTS OVER THE COMPLETE SEQUENCE MEASURED BY LOOP CLOSURES

Sequence	ORB-SLAM2 [24] (stereo)			LSD-SLAM [19]			DPO		
	scale	rot. [°]	pos. [%]	scale	rot. [°]	pos. [%]	scale	rot. [°]	pos. [%]
park. lot	1.02^{-1}	10.44	1.96	12.56^{-1}	10.01	33.83	1.07^{-1}	1.55	0.33
foodcourt	1.01^{-1}	3.77	0.75	2.60^{-1}	4.73	9.17	1.08	2.20	0.96
office	–	–	–	1.36^{-1}	7.86	4.67	1.16^{-1}	3.31	1.15

The table shows scale factor, rotational drift, and absolute position error from the beginning to the end of the sequences. The position error is given in percentage of the sequence length.

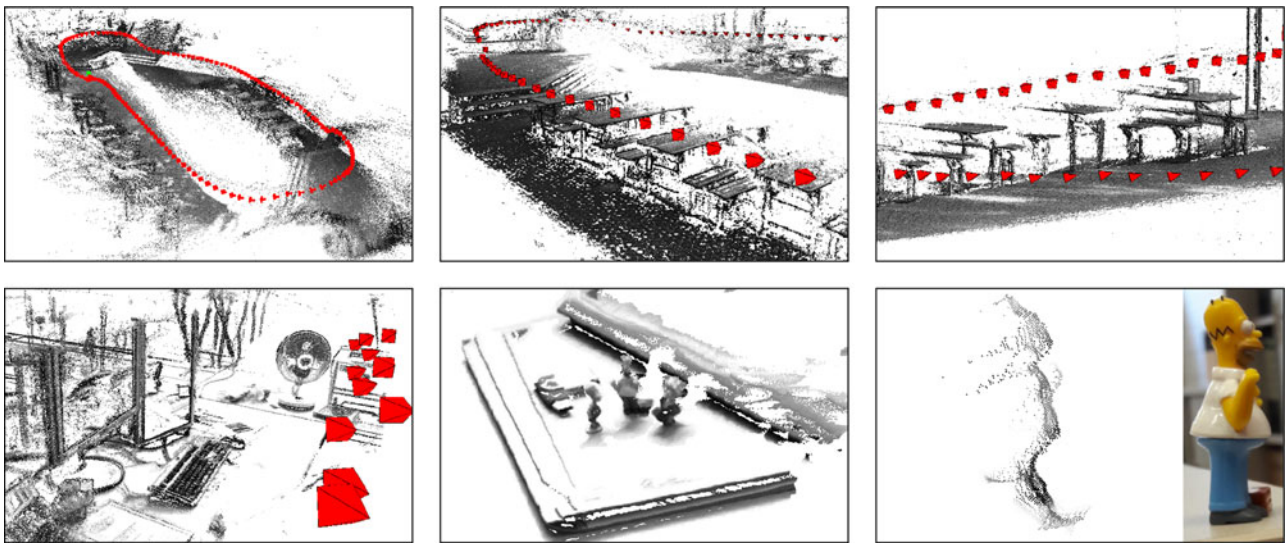


Fig. 8. Point cloud subsections created by DPO. Keyframe positions are marked in red. top: "foodcourt" sequence. bottom: "office" sequence. Homer figure (bottom, right) has a height of about 5cm.

64 images. The mean and the standard deviations of all intrinsic parameters are given for both methods in Table IV.

The parameters f_x, f_y, K_1 and K_2 need not to be calculated in our calibration approach. However, for comparison purposes we calculated these parameters (numbers in italic font) on the basis of our calibration results using the definitions given in [5].

As one can see from Table IV, the intrinsic parameters estimated with our calibration approach are an order of magnitude more robust than those of the method [5]. All estimates result in a relative standard deviation weigh less than one percent, except for the estimate of B. Probably the robustness of B could be improved by having more variation in the distance to the calibration target for the recorded images. A similar scattering as for B is received for K_2 since both parameters are highly correlated (see [5] for definition).

There is a somewhat large difference of the estimates of the intrinsic parameters f_x, f_y, K_1 and K_2 considering the two investigated methods. Our method estimates f_x and f_y to have the same value since we consider the pixels to be square. This is confirmed by the estimates of the method [5]. Since we do not have any absolute reference values we can not make any statement about the correctness of these values.

Both methods resulted in similar mean reprojection error, of 0.159 pixel for the method of Bok *et al.* [5] and 0.186 pixel for our method. This confirms the validity of both models.

B. Direct Plenoptic Odometry

Since there are no comparable plenoptic VO algorithms available we compare our method to state-of-the-art visual SLAM approaches: the monocular LSD-SLAM [19] and the stereo ORB-SLAM2 [24]. We ran our approach using the same camera as used for calibration (Section VII-A). The main focal length was chosen to $f_L = 16$ mm to achieve a suitable trade-off between FOV and scale awareness.

To compare the approaches, we mounted the plenoptic camera together with a stereo camera pair (image size: 1530×742 pixel, focal length: 705 pixel, baseline: 120 mm) on the same platform. We ran both camera systems with the same frame rate (15 fps) but without synchronization.

We present results for three different trajectories which were performed indoor as well as outdoor. For LSD-SLAM we reduced the image size to 640×480 pixel for which the implementation is optimized. For all sequences LSD-SLAM was initialized by the stereo depth map of the first frame, providing good starting conditions for this algorithm.

1) Quantitative Results: Since the frames of the sequences are not synchronized and we do not have a ground truth for the camera trajectory, we measure the overall drift based on a large loop closure. While we receive the loop closure data for the stereo camera sequences from the loop closure detection of

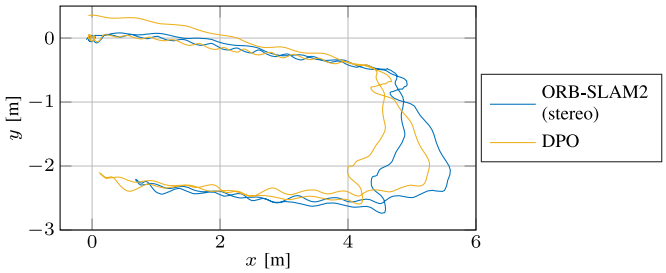


Fig. 9. Trajectory of "office" sequence estimated from ORB-SLAM2 (stereo) and DPO. The trajectory starts at coordinates (0,0,0), from there the cameras where moved on a U-shaped path around a group of tables and on a similar path back to the starting position.

ORB-SLAM2, we performed for DPO a 3D similarity transformation ($\mathrm{Sim}(3)$) based direct image alignment, as implemented in [19], between a representative keyframe at the beginning and the end of the sequence. The obtained co-registration from the beginning to the end of the sequences is used as ground truth data in the evaluation.

To obtain a fair comparison of all algorithms, for the evaluation large scale loop closure detection was disabled for ORB-SLAM2 and LSD-SLAM for all three sequences.

Table V presents the results for the three sequences. The table shows the scale factor, the rotational drift, and the absolute position error from the beginning to the end of the sequence.

For the "office" sequences ORB-SLAM2 results in a globally optimized trajectory since in the sequence a similar path was walked forward and backward (see Fig. 9). Therefore no drift could be measured. To avoid loop closures for LSD-SLAM on the "office" sequence we allowed map optimization only based on a neighborhood of ± 10 keyframes.

For the two large scale sequences ("parking lot" and "foodcourt") LSD-SLAM basically fails (scale drift $> 100\%$). Our DPO shows tracking accuracies comparable to ORB-SLAM2 (stereo) even though DPO performs only frame by frame tracking without any further optimization. DPO is only outperformed with respect to the scale drift, which is due to the much smaller static stereo baseline.

For all sequences ORB-SLAM2 (stereo) occasionally performed relocalization which means that tracking was lost temporarily and regained.

For the office sequence it might be that LSD-SLAM running on higher image resolution would result in less drifts. Nevertheless, for the two outdoor sequences ("parking lot" and "foodcourt") tracking of LSD-SLAM completely failed, which is not a result of the lower image resolution. The reason why LSD-SLAM fails is more because the camera is mainly moving towards viewing direction, which results in a biased $\mathrm{Sim}(3)$ estimation between consecutive keyframes and therefore in a continuously decreasing scale.

For both ORB-SLAM2 and LSD-SLAM we did not enforce real-time operation to obtain best possible results. Nevertheless, when rating the results given above, one has to consider that both algorithms are designed and implemented to run in real-time at the cost of poorer tracking and mapping performance. Our DPO in contrast currently exists only in a pure sequential, unoptimized implementation, which takes around 3 seconds to process a single frame. However, we showed already in previous publications [42] that depth estimation as a main part of DPO can run on frame rates > 30 fps on standard GPUs. Furthermore, direct image alignment on monocular images was also shown to be real-time capable [40]. Therefore, we are convinced that a real-time implementation of DPO, with similar performance as presented here, is feasible.

2) Qualitative Results: While on the outdoor trajectories ("parking lot" and "foodcourt") DPO has quite large absolute scale differences with respect to the stereo approach, on small scale it is able to estimate the scale from the light-field correctly. This is shown in Fig. 9, where the trajectories of DPO and ORB-SLAM2 (stereo) are shown for the "office" sequence.

Figs. 1 and 8 show subsections of the point clouds received from all three trajectories. Even though motion was performed mainly along the line of vision, DPO is still able to estimate accurate and detailed point clouds.

In close range DPO is able to estimate highly detailed point clouds with millimeter accuracy as can be seen from the silhouette of the Homer figure (Fig. 8, bottom, right), which was standing on the table in the "office" sequence and was recorded only from the front.

VIII. Summary and Conclusion

In this article we presented a complete framework to perform plenoptic camera based odometry. We developed a multiple view geometry for plenoptic cameras which enables tracking and mapping directly on the micro images at full sensor resolution. Based on this multiple view geometry we developed a calibration approach for plenoptic cameras which defines the projection from object space directly to the micro images on the sensor. By performing calibration based on a 3D calibration target, the resulting optimization problem is much better conditioned than in already existing methods. Our calibration method supplies a more robust estimate of the intrinsic camera model than a state-of-the-art calibration approach for MLA based light-field cameras.

The Direct Plenoptic Odometry (DPO) algorithm developed by us outperforms state-of-the-art monocular SLAM algorithms and is competitive to stereo approaches.

Although we do not have yet a real-time implementation of our DPO, we are convinced that an optimized implementation is able to run with high frame rates. Parts of the algorithm (depth estimation [42], direct image alignment [40]) have been shown to run at high frame rates.

Since there are no other plenoptic SLAM algorithms and respective datasets available yet, a direct comparison is not possible. Hence, this demands for a plenoptic odometry benchmark to compare future algorithms.

Acknowledgment

The authors would like to thank the anonymous reviewers and editors for their helpful comments.

REFERENCES

[1] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[2] R. Ng and P. Hanrahan, "Digital correction of lens aberrations in light field photography," *Proc. SPIE*, vol. 6342, 2006, Art. no. 63421E.

[3] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, "Modeling the calibration pipeline of the Lytro camera for high quality light-field image reconstruction," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 3280–3287.

[4] D. Dansereau, O. Pizarro, and S. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1027–1034.

[5] Y. Bok, H. G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 287–300, Feb. 2017.

[6] O. Johannsen, C. Heinze, B. Goldlücke, and C. Perwaß, "On the calibration of focused plenoptic cameras," in *Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications* (ser. Lecture Notes in Computer Science). Berlin, Germany: Springer, 2013, pp. 302–317.

[7] C. Heinze, S. Spyropoulos, S. Hussmann, and C. Perwaß, "Automated robust metric calibration algorithm for multifocus plenoptic cameras," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 5, pp. 1197–1205, May 2016.

[8] N. Zeller, F. Quint, and U. Stilla, "Calibration and accuracy analysis of a focused plenoptic camera," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. II-3, pp. 205–212, Sep. 2014.

[9] N. Zeller, F. Quint, and U. Stilla, "Depth estimation and camera calibration of a focused plenoptic camera for visual odometry," *ISPRS J. Photogramm. Remote Sens.*, vol. 118, pp. 83–100, 2016.

[10] N. Zeller, C.-A. Noury, F. Quint, C. Teulière, U. Stilla, and M. Dhôme, "Metric calibration of a focused plenoptic camera based on a 3D calibration target," in *Proc. ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, 2016, vol. III-3, pp. 449–456.

[11] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. IEEE ACM Int. Symp. Mixed Augmented Reality*, Nov. 2007, vol. 6, pp. 225–234.

[12] E. Eade and T. Drummond, "Edge landmarks in monocular SLAM," *Image Vision Comput.*, vol. 27, no. 5, pp. 588–596, 2009.

[13] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *Int. J. Robot. Res.*, vol. 32, no. 6, pp. 690–711, 2013.

[14] A. Concha and J. Civera, "Using superpixels in monocular SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 365–372.

[15] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[16] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 15–22.

[17] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2320–2327.

[18] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Proc. IEEE Int. Conf. Comput. Vision*, Dec. 2013, pp. 1449–1456.

[19] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 834–849.

[20] T. Schöps, J. Engel, and D. Cremers, "Semi-dense visual odometry for AR on a smartphone," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Sep. 2014, pp. 145–150.

[21] S. Izadi *et al.*, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 559–568.

[22] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 2100–2106.

[23] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc. Int. Conf. Intell. Robots Syst.*, Sep. 2015, pp. 1935–1942.

[24] R. Mur-Artal and J. D. Tards, "ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras," *ArXiv preprint arXiv:1610.06475*, 2016.

[25] D. Dansereau, I. Mahon, O. Pizarro, and S. Williams, "Plenoptic flow: Closed-form visual odometry for light field cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 4455–4462.

[26] F. Dong, S.-H. Ieng, X. Savatier, R. Etienne-Cummings, and R. Benosman, "Plenoptic cameras in real-time robotics," *Int. J. Robot. Res.*, vol. 32, no. 2, pp. 206–217, Feb. 2013.

[27] O. Johannsen, A. Sulc, and B. Goldlücke, "On linear structure from motion for light field cameras," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 720–728.

[28] T. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[29] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision—From Images to Geometric Models* (ser. Interdisciplinary Applied Mathematics), 1st ed. New York, NY, USA: Springer, 2004.

[30] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[31] R. Ng, "Digital light field photography," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, Jul. 2006.

[32] A. Lumsdaine and T. Georgiev, "Full resolution lightfield rendering," Adobe Systems, Inc., Mountain View, CA, USA, Tech. Rep. TR668, 2008.

[33] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. IEEE Int. Conf. Comput. Photography*, San Francisco, CA, USA, Apr. 2009, pp. 1–8.

[34] N. Zeller, F. Quint, and U. Stilla, "Establishing a probabilistic depth map from focused plenoptic cameras," in *Proc. Int. Conf. 3D Vision*, 2015, pp. 91–99.

[35] C. Hahne, A. Aggoun, S. Haxha, V. Velisavljevic, and J. C. J. Fernandez, "Baseline of virtual cameras acquired by a standard plenoptic camera setup," in *Proc. 3DTV-Conf.: The True Vision—Capture, Transmission Display 3D Video*, Jul. 2014, pp. 1–3.

[36] C. Perwaß and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," *Proc. SPIE*, vol. 8291, Jan. 2012, Art. no. 829108.

[37] D. C. Brown, "Decentering distortion of lenses," *Photogramm. Eng.*, vol. 32, no. 3, pp. 444–462, May 1966.

[38] S. Agarwal *et al.*, "Ceres solver," 2012. [Online]. Available: http://ceres-solver.org

[39] N. Zeller, F. Quint, and U. Stilla, "Filtering probabilistic depth maps received from a focused plenoptic camera," in *Proc. 2nd BW-CAR Symp. Inf. Commun. Syst.*, 2015, vol. 2, pp. 7–12.

[40] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 3748–3754.

[41] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[42] R. Vasko, N. Zeller, F. Quint, and U. Stilla, " A real-time depth estimation approach for a focused plenoptic camera," in *Advances in Visual Computing* (ser. Lecture Notes in Computer Science), vol. 9475. New York, NY, USA: Springer, 2015, pp. 70–80.

Author's photographs and biographies not available at the time of publication.