ISPRS Journal of Photogrammetry and Remote Sensing 118 (2016) 83–100

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Depth estimation and camera calibration of a focused plenoptic camera for visual odometry

Niclas Zeller [a,b,*], Franz Quint [a], Uwe Stilla [b]

[a] Faculty of Electrical Engineering and Information Technology, Karlsruhe University of Applied Sciences, Moltkestraße 30, 76133 Karlsruhe, Germany
[b] Department of Photogrammetry and Remote Sensing, Technische Universität München, Arcisstraße 21, 80333 München, Germany

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form 15 March 2016
Accepted 27 April 2016

Keywords:
Accuracy analysis
Camera calibration
Focused plenoptic camera
Light-field
Probabilistic depth map
Visual odometry

ABSTRACT

This paper presents new and improved methods of depth estimation and camera calibration for visual odometry with a focused plenoptic camera.

For depth estimation we adapt an algorithm previously used in structure-from-motion approaches to work with images of a focused plenoptic camera. In the raw image of a plenoptic camera, scene patches are recorded in several micro-images under slightly different angles. This leads to a multi-view stereo-problem. To reduce the complexity, we divide this into multiple binocular stereo problems. For each pixel with sufficient gradient we estimate a virtual (uncalibrated) depth based on local intensity error minimization. The estimated depth is characterized by the variance of the estimate and is subsequently updated with the estimates from other micro-images. Updating is performed in a Kalman-like fashion. The result of depth estimation in a single image of the plenoptic camera is a probabilistic depth map, where each depth pixel consists of an estimated virtual depth and a corresponding variance.

Since the resulting image of the plenoptic camera contains two plains: the optical image and the depth map, camera calibration is divided into two separate sub-problems. The optical path is calibrated based on a traditional calibration method. For calibrating the depth map we introduce two novel model based methods, which define the relation of the virtual depth, which has been estimated based on the light-field image, and the metric object distance. These two methods are compared to a well known curve fitting approach. Both model based methods show significant advantages compared to the curve fitting method.

For visual odometry we fuse the probabilistic depth map gained from one shot of the plenoptic camera with the depth data gained by finding stereo correspondences between subsequent synthesized intensity images of the plenoptic camera. These images can be synthesized totally focused and thus finding stereo correspondences is enhanced. In contrast to monocular visual odometry approaches, due to the calibration of the individual depth maps, the scale of the scene can be observed. Furthermore, due to the light-field information better tracking capabilities compared to the monocular case can be expected.

As result, the depth information gained by the plenoptic camera based visual odometry algorithm proposed in this paper has superior accuracy and reliability compared to the depth estimated from a single light-field image.

1. Introduction

The concept of a plenoptic camera has been developed more than hundred years ago by Ives (1903) and Lippmann (1908). However, only for the last few years the existing graphic processor units (GPUs) are capable to evaluate the recordings of a plenoptic camera with acceptable frame rates ($\geqslant$25 fps).

To gather the whole 4D light-field (hence the name "plenoptic"), a microlens array (MLA) is placed in front of the sensor. Today there exist basically two main concepts of MLA based plenoptic cameras: The unfocused plenoptic camera proposed by Adelson and Wang (1992) and developed further by Ng (2006) and the focused plenoptic camera, which was described for the first time by Lumsdaine and Georgiev (2008). Compared to the unfocused plenoptic camera, the focused plenoptic camera has a higher spatial resolution but lower angular resolution. This high spatial resolution is especially beneficial for estimating depth out

* Corresponding author.
E-mail addresses: niclas.zeller@hs-karlsruhe.de (N. Zeller), franz.quint@hs-karlsruhe.de (F. Quint), stilla@tum.de (U. Stilla).

http://dx.doi.org/10.1016/j.isprsjprs.2016.04.010
0924-2716/

of the recorded raw image (Perwaß and Wietzke, 2012). In our research we are using a focused plenoptic camera.

The accuracy of the depth information gathered by a focused plenoptic camera is rather low for a distance of a few meters compared to other depth sensors, like Time-of-Flight (TOF) cameras or stereo camera systems with a large baseline, at least at a comparable field of view (FOV). Besides, the depth accuracy of a focused plenoptic camera strongly decays when reducing the focal length. Thus, a trade-of between wide FOV and acceptable accuracy has to be found.

On the other hand as shown by Perwaß and Wietzke (2012), the plenoptic camera offers a much larger depth of field (DOF) compared to a monocular camera at the same aperture. Thus, a plenoptic camera has a much shorter close range limit than e.g. a stereo camera system.

Another plus for plenoptic cameras are their small dimensions, which are similar to those of a conventional camera. In future there will also be miniaturized light-field sensors available, which can be assembled in smartphones (Venkataraman et al., 2013).

In many navigation applications such small sensors are profitable, for example on unmanned aerial vehicles (UAVs), where space and weight is limited. But also for indoor navigation or blind people assistance, where bulky sensors can be annoying, such small and light sensors are beneficial.

For this kind of applications today mostly monocular visual odometry (or Simultaneous Localization and Mapping (SLAM)) systems are used, which gain depth information from camera motion. However, such monocular systems come with some drawbacks. One drawback of a monocular visual odometry system is its scale ambiguity. Thus, especially in navigation applications additional sensors are needed to gather metric dimensions. Another disadvantage of the monocular system is that no depth is obtained without any motion of the camera or for rotations around the camera's optical center.

Thus, a plenoptic camera seems to be a good compromise between a monocular and a stereo camera for a visual odometry system. Since for a plenoptic camera rough depth information is available for each single frame, it is to be expected that tracking will become much more robust compared to a monocular system.

Fig. 1 shows two typical scenarios for indoor navigation recorded by a plenoptic camera. Here, far as well as very close objects with less than one meter distance to the camera are present in the same scene. In such scenes a plenoptic camera benefits from its large DOF. Even though the scene has a high variation in depth the camera is able to record the whole scene in focus.

The presented work unifies and extends the previous publications (Zeller et al., 2014, 2015a,b). Thereby, this paper provides the complete work-flow of a focused plenoptic camera based visual odometry. We firstly present the concept of the focused plenoptic camera and how depth can be estimated in principle out of the recorded light-field (Section 2). Additionally, we derive the theoretically achievable depth accuracy of the camera (Section 3). Afterwards, we propose a new depth estimation algorithm which estimates a probabilistic depth map from a single recording of a focused plenoptic camera (Section 4). The probabilistic depth map will be beneficial for the visual odometry system. Like any camera system that is used for photogrammetric purposes, a plenoptic camera has to be calibrated. We present a complete framework to calibrate a focused plenoptic camera, especially for an object distance range of several meters (Section 5). For this we develop three different depth models and compare them to each other. Finally we incorporate the depth estimation as well as the camera calibration into a focused plenoptic camera based visual odometry system (Section 6). All proposed methods are extensively evaluated (Section 7).

1.1. Related work

1.1.1. Depth estimation

For the last years various algorithms for depth estimation based on the recordings of plenoptic cameras or other light-field representations have been developed. First methods were published even more than 20 years ago (Adelson and Wang, 1992).

Since light-field based depth estimation represents a multi-dimensional optimization problem, always a trade-off between low complexity and high accuracy or consistency has to be chosen. Wanner and Goldluecke (2012, 2014) for instance present a globally consistent depth labeling which is performed directly on the 4D light-field representation and results in a dense depth map. Jeon et al. (2015) make use of the phase-shift theorem of the Fourier transform to calculate a dense, light-field based disparity map with sub-pixel accuracy, while Heber and Pock (2014) use principal component analysis to find the optimum depth map. Some approaches make use of geometric structures like 3D line segments (Yu et al., 2013) to improve the estimate and to reduce complexity. Tosic and Berkner (2014) present a so called scale-depth space which provides a coarse depth map for uniform regions and a fine one for textured regions. Other methods reduce complexity by the use of local instead of global constraints and thus result in a sparse depth map. Such sparse maps supply depth only for textured

(a) (b)

Fig. 1. Two scenarios typical for indoor navigation recorded by a focused plenoptic camera. Here, far as well as very close objects with less than one meter distance to the camera are present in the same scene. Due to the large DOF of a plenoptic camera such scenes with a high variation in depth still can be recorded completely in focus by the plenoptic camera.

regions (Bishop and Favaro, 2011; Perwaß and Wietzke, 2012). The methods presented by Tao et al. (2013) and Kim et al. (2014) additionally make use of the focus cues which are supplied by a plenoptic camera.

1.1.2. Plenoptic camera calibration

There exist already different publications which describe the calibration of an MLA based plenoptic camera. While Dansereau et al. (2013) describe in general form the calibration of MLA based plenoptic cameras, Johannsen et al. (2013) present the calibration of a focused plenoptic camera for object distances up to about 50 cm. We refer in this paper to the same plenoptic camera (Perwaß and Wietzke, 2012) but our approach can handle by orders of magnitude larger depth ranges. Furthermore, we are able to use a simple setup to calibrate the camera.

1.1.3. Visual odometry

To accomplish their task, visual odometry or SLAM systems can use feature-based and direct methods. Some systems are based on depth or stereo image sensors, while other are monocular.

In feature-based visual odometry, features are extracted from the recorded 2D images by using some feature detector (Klein and Murray, 2007; Eade and Drummond, 2009; Li and Mourikis, 2013; Concha and Civera, 2014). The features are matched between the corresponding images. Based on the feature correspondences, the camera position and the 3D feature point coordinates are estimated. From a feature-based method only a sparse point cloud is received.

Direct methods, like the one presented by Forster et al. (2014), perform tracking and mapping directly on the recorded images. Tracking becomes much more robust since all image data is used. As presented by Newcombe et al. (2011), direct methods can be used to estimate a dense depth map. Such direct, dense methods are very complex. The complexity can be reduced by performing semi-dense direct tracking and mapping algorithms (Engel et al., 2013; Engel et al., 2014). Semi-dense means, that only image regions of high contrast are considered for tracking and mapping and all homogeneous regions are neglected. These semi-dense methods are capable to run in real-time on today's standard central processing units (CPUs) or even on smartphones (Schöps et al., 2014).

The use of multiple cameras or depth sensors strongly simplifies the visual odometry problem. Here depth information is already received without motion. Besides, the scale of the scene is received directly from the recorded images without using any additional sensors (Akbarzadeh et al., 2006; Izadi et al., 2011; Dansereau et al., 2011; Kerl et al., 2013).

2. The focused plenoptic camera

In this section we will present the concept of the focused plenoptic camera, which is used in our research and we will establish the equations to retrieve depth information from the recorded light-field.

As presented by Lumsdaine and Georgiev (2009) a focused plenoptic camera can be realized in two different configurations, as shown in Fig. 2: the Keplerian configuration and the Galilean configuration.

In the Keplerian configuration an MLA and the sensor are placed behind the focused image which is created by the main lens (Fig. 2 (a)). Here, the focal length of the micro lenses is chosen such that multiple focused sub-images (micro images) of the main lens image occur on the image sensor.

In the Galilean configuration MLA and sensor are placed in front of the focused image which would be created by the main lens behind the sensor (Fig. 2(b)). Subsequently we will call this image behind the sensor the virtual image. Similar to the Keplerian configuration, the focal length of the micro lenses is chosen such that multiple sub-images of the virtual image occur focused on the image sensor.

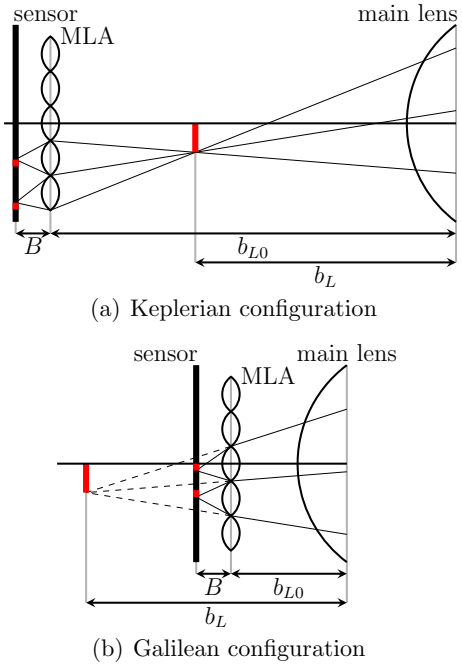


Fig. 2. Different configurations of a focused plenoptic camera. (a) Keplerian configuration: the MLA is placed behind the image which is created by the main lens. (b) Galilean configuration: the MLA is placed in front of the "virtual" image which would be created by the main lens.

The camera which we are using in our research is of Galilean configuration and is from the manufacturer Raytrix. While a plenoptic camera has already a larger DOF than a monocular camera at the same main lens aperture (Georgiev and Lumsdaine, 2009; Perwaß and Wietzke, 2012), in a Raytrix camera the DOF is further increased by using an interlaced MLA in a hexagonal arrangement. This MLA consists of three different micro lens types. Each type has a different focal length and thus focuses a different virtual image distance on the sensor. The DOFs of the three micro lens types are chosen such that they are just adjacent to each other. Thus, the effective DOF of the camera is increased compared to an MLA with only one type of micro lenses (Perwaß and Wietzke, 2012).

In the following we will only discuss a focused plenoptic camera which relies on the Galilean configuration. Nevertheless, for the Keplerian configuration similar relations can be derived.

2.1. Path of rays of a Galilean focused plenoptic camera

If we consider the main lens to be an ideal thin lens, the relationship between the object distance a_L of an object point and the image distance b_L of the corresponding image point is defined by the thin lens equation given in Eq. (1).

$$\frac{1}{f_L} = \frac{1}{a_L} + \frac{1}{b_L} \tag{1}$$

Here f_L is the main lens focal length. Thus, if the image distance b_L of an image point is known, the object distance a_L of the corresponding object point can be calculated.

The easiest way to understand the principle of a plenoptic camera is to consider only the path of rays inside the camera, as shown

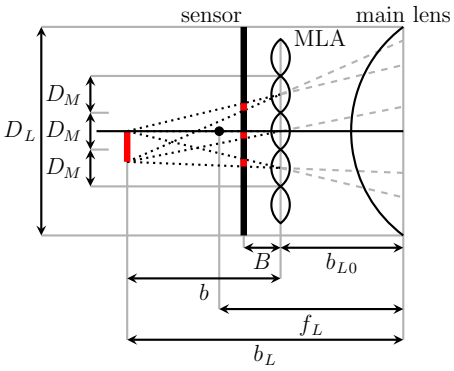


Fig. 3. Optical path inside a focused plenoptic camera based on the Galilean configuration. The MLA and the image sensor lie in front of the "virtual" image which would be created by the main lens. A virtual image point in distance b behind the MLA results in multiple focused micro images on the sensor.

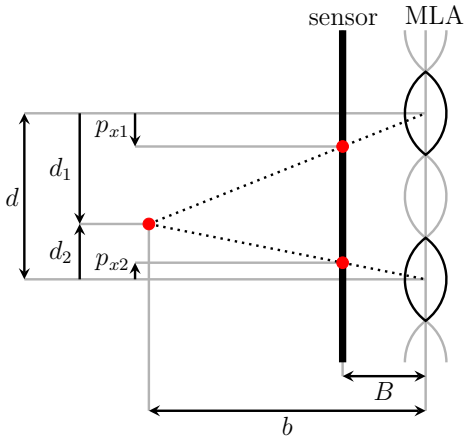


Fig. 4. Depth estimation in a focused plenoptic camera based on the Galilean configuration. The distance b between a virtual image point and the MLA can be calculated based on its projection in two or more micro images.

in Fig. 3. For the following descriptions the MLA is assumed to be a pinhole grid, which simplifies the path of rays. Thus, each pixel on the image sensor can be considered as the endpoint of the central ray through the corresponding micro lens. The image on the sensor can be interpreted as a 4D (non-uniformly) sampled representation of the light-field L inside the camera, as follows (Gortler et al., 1996):

$$L \sim f(x, y, \phi, \theta) \tag{2}$$

where x and y define the position of the corresponding micro lens center and ϕ and θ the incident angle on the MLA plane. By projecting the sampled light-field through the main lens, the corresponding 4D light-field in object space is received.

2.2. Retrieving depth from the sampled light-field

In general, a virtual image point is projected to multiple micro images. For instance Fig. 3 shows how the virtual image is projected by the three middle micro lenses onto different pixels of the sensor. If it is known which pixels on the sensor correspond to the same virtual image point, the distance b between MLA and virtual image can be calculated by triangulation.

To derive how the distance b can be calculated, Fig. 4 shows exemplary the triangulation for a virtual image point based on the corresponding points in two micro images.

In Fig. 4 p_{xi} (for $i \in \{1,2\}$) define the distances of the points in the micro images with respect to the principal points of their micro images. Similarly, d_i (for $i \in \{1,2\}$) define the distances of the respective principal points to the orthogonal projection of the virtual image point on the MLA. All distances p_{xi}, as well as d_i are defined as signed values. Distances with an upwards pointing arrow in Fig. 4 have positive values and those with a downwards pointing arrow have negative values. Triangles which have equal angles are similar and the following relations hold:

$$\frac{p_{xi}}{B} = \frac{d_i}{b} \quad \longrightarrow \quad p_{xi} = \frac{d_i \cdot B}{b} \quad \text{for} \quad i \in \{1,2\} \tag{3}$$

The baseline distance d between the two micro lenses can be calculated as given in Eq. (4).

$$d = d_2 - d_1 \tag{4}$$

We define the disparity p_x of the virtual image point as the difference between p_{x2} and p_{x1}, and yield the relation given in Eq. (5):

$$p_x = p_{x2} - p_{x1} = \frac{(d_2 - d_1) \cdot B}{b} = \frac{d \cdot B}{b} \tag{5}$$

After rearranging Eq. (5), the distance b between a virtual image point and the MLA can be described as a function of the baseline distance d, the distance B between MLA and sensor, and the disparity p_x, as given in Eq. (6).

$$b = \frac{d \cdot B}{p_x} \tag{6}$$

A virtual image point occurs in more or less micro images depending on its distance b to the MLA. Thus, the length of the longest baseline d, which can be used for triangulation, changes. It can be defined as a multiple of the micro lens diameter $d = k \cdot D_M$ ($k \geqslant 1$). Here, k is not mandatory an integer, due to e.g. the 2D hexagonal arrangement of the micro lenses on the MLA.

The baseline distance d and the disparity p_x are both defined in pixels and can be measured from the recorded micro lens images,[1] while the distance B between MLA and sensor is a metric dimension which cannot be measured precisely. Thus, the distance b is estimated relatively to the distance B. This relative distance, which is free of any unit is called virtual depth and will be denoted by v in the following.

$$v = \frac{b}{B} = \frac{d}{p_x} \tag{7}$$

To retrieve the real depth, i.e. the distance a_L in object space between an observed point and the camera, one has to estimate the relation between the virtual depth v and the image distance b_L (which relies on B and b_{L0}) in a calibration process. Then, one can use the thin lens equation (Eq. (1)) to calculate the object distance a_L. We will derive in Section 3, how uncertainties in the estimation of the disparity propagate to the object distance.

2.3. Image synthesis

Within the DOF of the plenoptic camera it is known that each virtual image point $\boldsymbol{x}_V = (x_V, y_V)^T$ occurs focused in at least one micro image. Thus, based on the virtual depth $v(\boldsymbol{x}_V)$ a so called totally focused image $I(\boldsymbol{x}_V)$ can be synthesized. Here the intensity value is calculated as the average over all corresponding focused micro image points. For further details on the image synthesis we refer to Perwaß and Wietzke (2012).

[1] When we refer to the micro lens images, debayering and white balancing is considered to be already performed.

3. Derivation of theoretically achievable depth accuracy

Based on the rules known from the theory of propagation of uncertainty one can derive how the uncertainty σ_{p_x} of the estimated disparity p_x will effect the uncertainty σ_{a_L} of the object distance a_L. From the derivative of v with respect to the measured disparity p_x the standard deviation of the virtual depth σ_v can be approximated as given in Eq. (8).

$$\sigma_v \approx \left|\frac{\partial v}{\partial p_x}\right| \cdot \sigma_{p_x} = \frac{d}{p_x^2} \cdot \sigma_{p_x} = \frac{v^2}{d} \cdot \sigma_{p_x} \tag{8}$$

Eq. (8) shows that the accuracy of the virtual depth decays proportional to v^2 if the baseline distance d is constant. On the other hand, Eq. (8) reveals, that long baselines d are beneficial for a good accuracy of the virtual depth. However, Fig. 3 shows, that long baselines can only be used for points which have a high virtual depth.

As stated in the prior section, the baseline distance d is a multiple of the micro lens diameter $D_M : d = k \cdot D_M$ ($k \geqslant 1$). When the virtual depth increases (for objects moving close to the camera), one can switch to longer baselines d, resulting in a discontinuous dependency of σ_v from σ_{p_x}. This finally leads to a discontinuous dependency of the depth accuracy as function of the object distance a_L.

The relationship between the image distance b_L and the virtual depth v is defined by the linear function given in Eq. (9).

$$b_L = b + b_{L0} = v \cdot B + b_{L0} \tag{9}$$

Here b_{L0} is the unknown but constant distance between main lens and MLA. Using the thin lens equation (Eq. (1)) one can finally express the object distance a_L as function of the virtual depth v. If the derivative of a_L with respect to b_L is calculated, the standard deviation of the object distance σ_{a_L} can be approximated as given in Eq. (10).

$$\sigma_{a_L} \approx \left|\frac{\partial a_L}{\partial b_L}\right| \cdot \sigma_{b_L} = \frac{f_L^2}{(b_L - f_L)^2} \cdot \sigma_{b_L} = \frac{(a_L - f_L)^2}{f_L^2} \cdot \sigma_{b_L} = \frac{(a_L - f_L)^2}{f_L^2} \cdot B \cdot \sigma_v \tag{10}$$

For object distances which are much higher than the focal length of the main lens f_L the approximation in Eq. (10) can be further simplified as given in Eq. (11). From Eq. (11) one can see, that for a constant object distance a_L the depth accuracy increases proportional to f_L^2.

$$\sigma_{a_L} \approx \frac{a_L^2}{f_L^2} \cdot B \cdot \sigma_v \quad \text{for} \quad a_L \gg f_L \tag{11}$$

From Eqs. (8)–(10) one receives σ_{a_L} with respect to σ_{p_x}:

$$\sigma_{a_L} = \frac{f_L^2}{(b_L - f_L)^2} \cdot B \cdot \frac{v^2}{d} \cdot \sigma_{p_x} = \frac{f_L^2}{(v \cdot B + b_{L0} - f_L)^2} \cdot B \cdot \frac{v^2}{d} \cdot \sigma_{p_x} \tag{12}$$

Interesting cognition is received when we assume the MLA to lie in the focal plane ($f_L = b_{L0}$):

$$\sigma_{a_L} = \frac{f_L^2}{d \cdot B} \cdot \sigma_{p_x} = \frac{f_L^2}{k \cdot D_M \cdot B} \cdot \sigma_{p_x} \tag{13}$$

From Eq. (13) one can see, that the camera supplies a constant depth accuracy σ_{a_L} for a certain baseline distance. Nevertheless, in this setup the camera will not be able to capture object distances a_L up to infinity and therefore is not suitable for visual odometry. For a distance $b_{L0} > f_L$, σ_{a_L} even decreases with increasing object distance a_L. Nevertheless, in that case the operating range of the camera is limited to a very short range close to the camera.

To be able to reconstruct a focused image from the recorded micro images of the focused plenoptic camera, it has to be assured that each point occurs focused in at least one micro image. As described by Perwaß and Wietzke (2012), for a focused plenoptic camera with tree different micro lens types in a hexagonally arranged MLA (as it is for Raytrix cameras) this is the case for all points which have a virtual depth $v \geq 2$. Hence, the plane at $v = 2$ is also called total covering plane (TCP). Therefore, the shortest image distance $b_{L\,\min}$, which refers the longest object distance $a_{L\,\max}$, has to satisfy the following relation:

$$b_{L\,\min} \geqslant 2 \cdot B + b_{L0} \tag{14}$$

Thus, to be able to synthesize images of scenes with object distances up to infinity ($a_{L\max} \to \infty$, hence $b_{L\min} \to f_L$) one has to assure that $f_L \geqslant 2 \cdot B + b_{L0}$.

4. Virtual depth estimation

Virtual depth estimation is basically about finding corresponding points in the micro images and solving Eq. (7). We consider the virtual depth estimation as a multi-view stereo problem since each virtual image point occurs in multiple micro images. It simplifies in the sense that all micro lenses have the same orientation by construction and thus the micro images are already rectified. Besides, since the micro images have a very small number of pixels ($\sim$ 23 pixel diameter) and a relatively narrow FOV ($\sim 25^\circ$), distortions caused by the micro lenses are neglected.

However, finding simultaneously correspondences across multiple micro images leads to a computationally demanding search problem. Furthermore, since the micro images have a very small diameter, feature extraction and matching seems to be error prone. We follow a different approach which is based on multiple depth observation received from different micro image pairs. Instead of feature matching we determine pixel correspondences by intensity error minimization along the epipolar line. For each depth observation an uncertainty measure is defined and thus, a probabilistic virtual depth map is established. This approach is similar to the one of Engel et al. (2013) where it is used to gain depth in a monocular visual odometry approach.

4.1. Probabilistic virtual depth

We define the inverse virtual depth $z = v^{-1}$, which is obtained from Eq. (7). The inverse virtual depth z is proportional to the estimated disparity p_x, as given in Eq. (15).

$$z = \frac{1}{v} = \frac{p_x}{d} \tag{15}$$

Since we determine pixel correspondences by matching pixel intensities, the sensor noise will be the main error source which affects the disparity estimation and thus the inverse virtual depth z. We neglect for instance misalignment of the MLA with respect to the image sensor or offsets on the micro lens centers. However, since these errors are not stochastic but have manufacturing reasons, they could be eliminated in a calibration process. Furthermore, as one can see from Eq. (15), the estimate of z relies only on the baseline distance d and the disparity p_x which result both as differences of absolute 2D positions in pixel coordinates (see Eq. (5)). Thus, at least within a local region, the estimate of z is invariant to alignment errors on the MLA.

The sensor noise is usually modeled as additive white Gaussian noise (AWGN). Since pixel correspondences are estimated based on intensity values, the disparity p_x and thus the estimated inverse virtual depth z can also be considered as Gaussian distributed. This projection will be derived mathematically in Section 4.3.2.

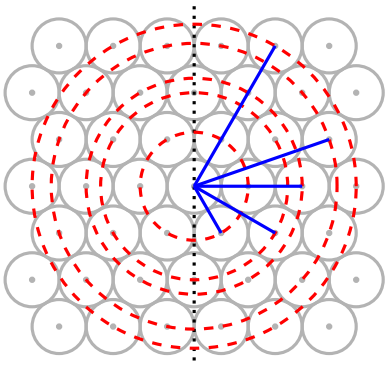


Fig. 5. Five shortest baseline distances in a hexagonal micro lens grid. For one micro lens, stereo matching is only performed with neighbors for which the baseline angle ϕ is in the range $-90^\circ \leqslant \phi < 90^\circ$.

In the following we will denote the inverse virtual depth hypothesis of a pixel by the random variable $Z \sim \mathcal{N}(z, \sigma_z^2)$ defined by the density function $f_Z(x)$, as given in Eq. (16).

$$f_Z(x) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(x-z)^2}{2\sigma_z^2}} \tag{16}$$

4.2. Graph of baselines

For stereo matching we define a graph of baselines. This graph defines which micro images are matched to each other. Each baseline in the graph is given by its length d and its 2D orientation on the MLA plane $\boldsymbol{e}_p = (e_{px}, e_{py})^T$. Since the micro images are all rectified, the orientation vector of the baseline is equivalent to that of the epipolar line. Thus, $\boldsymbol{e}_p$ defines the epipolar line for each pixel of the micro lens pair. In the following we will always consider $\boldsymbol{e}_p$ to be normed to unity ($\|\boldsymbol{e}_p\| = 1$ pixel).

In the graph the baselines are sorted in ascending order with respect to their length. This is also the order in which stereo matching will be performed. Matching is performed in that order since for short baselines it is more likely to find a unique match. For long baselines it is more likely to find ambiguous matches, but on the other hand the depth estimate will be more accurate. Thus, the matching result for short baselines can be used as prior knowledge for micro image pairs which are connected by a longer baseline.

Since stereo matching is performed for each micro lens separately, matching is only performed with respect to micro images right to the micro image of interest. Thus, only baselines or epipolar lines with an angle $-90^\circ \leqslant \phi < 90^\circ$ are considered.

Fig. 5 shows the five shortest baseline distances in a hexagonal MLA grid. Here the red[2] dashed circles represent the respective baseline distance around the micro lens of interest. The solid blue lines show one example baseline for each distance, while only baselines right of the dotted line are used for stereo matching. The epipolar line $\boldsymbol{e}_p$ is defined such that it points away from the reference micro lens.

4.3. Virtual depth observation

The inverse virtual depth estimation is performed for each pixel $\boldsymbol{x}_R = (x_R, y_R)^T$ in the sensor image. Prior to the estimation the vignetting resulting from the micro lenses is corrected by dividing the raw image by a priorly recorded white image. Additionally, the raw image is converted into a gray-scale image based on which the inverse virtual depth estimation is performed.

As already mentioned, the depth observation is performed starting from the shortest baseline up to the largest possible baseline. Based on each new observation, the inverse depth hypothesis of a raw image pixel $Z(\boldsymbol{x}_R)$ is updated and thus becomes more reliable.

To reduce computational effort, for each baseline it is checked first, if the pixel under consideration $\boldsymbol{x}_R$ has sufficient contrast along the epipolar line, as defined in Eq. (17).

$$|\boldsymbol{g}_I(\boldsymbol{x}_R)^T \boldsymbol{e}_p| \geqslant T_H \tag{17}$$

Here $\boldsymbol{g}_I(\boldsymbol{x}_R)$ represents the intensity gradient vector at the coordinate $\boldsymbol{x}_R$ and T_H some predefined threshold.

4.3.1. Stereo-matching

To find the pixel in a certain micro image which corresponds to the pixel of interest $\boldsymbol{x}_R$ we search for the minimum intensity error along the epipolar line in the corresponding micro image.

If there was no inverse virtual depth observation obtained yet, for the pixel of interest $\boldsymbol{x}_R$ an exhaustive search along the epipolar line has to be performed. For that case the search range is limited on one end by the micro lens border and on the other end by the coordinates of $\boldsymbol{x}_R$ with respect to the micro lens center. A pixel on the micro lens border results in the maximum observable disparity p_x and thus in the minimum observable virtual depth v. A pixel at the same coordinates as the pixel of interest in the corresponding micro image equals a disparity $p_x = 0$ and thus a virtual depth $v = \infty$. Of course the initial search range could be limited in advance to a certain virtual depth range $v_{\min} \leqslant v \leqslant v_{\max}$.

If there already exists an inverse virtual depth hypothesis $Z(\boldsymbol{x}_R)$, the search range can be limited to $z(\boldsymbol{x}_R) \pm n\sigma_z(\boldsymbol{x}_R)$, where n is usually chosen to be $n = 2$.

In the following we define the search range along the epipolar line as given in Eq. (18).

$$\boldsymbol{x}_R^s(p_x) = \boldsymbol{x}_{R0}^s - p_x \cdot \boldsymbol{e}_p \tag{18}$$

Here $\boldsymbol{x}_{R0}^s$ is defined as the coordinate of a point on the epipolar line at the disparity $p_x = 0$, as given in Eq. (19).

$$\boldsymbol{x}_{R0}^s = \boldsymbol{x}_R + d \cdot \boldsymbol{e}_p \tag{19}$$

Within the search range we calculate the sum of the squared intensity error e_{ISS} over a 1-dimensional pixel patch $(1 \times N)$ along the epipolar line, as defined in Eq. (20).

$$\begin{aligned} e_{ISS}(p_x) &= \sum_{k=-\frac{N-1}{2}}^{\frac{N-1}{2}} \left[I(\boldsymbol{x}_R + k\boldsymbol{e}_p) - I(\boldsymbol{x}_R^s(p_x) + k\boldsymbol{e}_p) \right]^2 \\ &= \sum_{k=-\frac{N-1}{2}}^{\frac{N-1}{2}} e_I(p_x + k)^2 \end{aligned} \tag{20}$$

The best match is the disparity p_x which minimizes $e_{ISS}(p_x)$. For the experiments presented in Section 7.1 we set $N = 5$. In the following we refer to the estimated disparity by $\hat{p}_x$, which defines the corresponding pixel coordinate $\boldsymbol{x}_R^s(\hat{p}_x)$. It is important to emphasize that p_x is estimated with sub-pixel accuracy by interpolating linearly between the samples of the intensity image $I(\boldsymbol{x}_R)$ and therefore is not restricted to any regular grid.

4.3.2. Observation uncertainty

The sensor noise n_I is the main error source which effects the estimated disparity $\hat{p}_x$ and thus the inverse virtual depth observation.

In our approach the variance of the raw image noise σ_N^2 is considered to be the same for each pixel $\boldsymbol{x}_R$. Actually, this holds only in

[2] For interpretation of color in Figs. 5, 6 and 8, the reader is referred to the web version of this article.

first assumption since, by correcting the vignetting, the noise at the micro image borders is amplified.

In the following it will be derived how σ_N^2 effects the disparity estimation. Therefore, we formulate the stereo matching by the minimization problem given in Eq. (21), where the estimated disparity $\hat{p}_x$ is the one which minimizes the squared intensity error $e_I(p_x)^2$. For a temporary simplification of the expressions, we omit the sum over a number of pixels, as defined for $e_{ISS}(p_x)$ in Eq. (20).

$$\hat{p}_x = \min_{p_x}\left[e_I(p_x)^2\right] = \min_{p_x}\left[\left(I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x))\right)^2\right] \tag{21}$$

To find the minimum, we calculate the first derivation of the error with respect to p_x and set it to zero. It results Eq. (23) as long as $g_I(p_x) \neq 0$ holds.

$$\begin{aligned}\frac{\partial e_I(p_x)^2}{\partial p_x} &= \frac{\partial\left[I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x))\right]^2}{\partial p_x} \\ &= 2\left[I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x))\right]\cdot\left[-g_I(p_x)\right]\end{aligned} \tag{22}$$

$$0 \stackrel{!}{=} I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x)) \tag{23}$$

Here, the intensity gradient along the epipolar line $g_I(p_x)$ is defined as follows:

$$g_I(p_x) = g_I(\boldsymbol{x}_R^s(p_x)) = \frac{\partial I(\boldsymbol{x}_{R0}^s - p_x\boldsymbol{e}_p)}{\partial p_x} \tag{24}$$

After approximating Eq. (23) by its first order Taylor-series it can be solved for p_x as given in Eq. (25).

$$\hat{p}_x = \frac{I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_{x0}))}{g_I(\boldsymbol{x}_R^s(p_{x0}))} + p_{x0} \tag{25}$$

If we now consider $I(\boldsymbol{x}_R)$ in Eq. (25) as a Gaussian distributed random variable, the variance $\sigma_{p_x}^2$ of the disparity p_x can be derived as given in Eq. (26).

$$\sigma_{p_x}^2 = \frac{\mathrm{Var}\{I(\boldsymbol{x}_R)\} + \mathrm{Var}\{I(\boldsymbol{x}_R^s(p_{x0}))\}}{g_I(\boldsymbol{x}_R^s(p_{x0}))^2} = \frac{2\sigma_N^2}{g_I(\boldsymbol{x}_R^s(p_{x0}))^2} \tag{26}$$

Fig. 6 illustrates how the gradient g_I effects the estimation of p_x. The blue line represents the tangent at the disparity p_{x0} at which the intensity values are projected onto the disparities.

Beside the stochastic error which results from the sensor noise, there will also be a systematic error present for a Raytrix camera when performing stereo matching in neighboring micro images. This systematic error occurs since neighboring micro images have different focal lengths. Thus, for the same virtual image region some micro images will be in focus while others will be blurred. Hence, beside the variance $\sigma_{p_x}^2$ we define a second error source which models differences in the projection process. In this so called focus uncertainty we take into account that a small intensity error e_{ISS} very likely gives a more reliable disparity estimate than a large

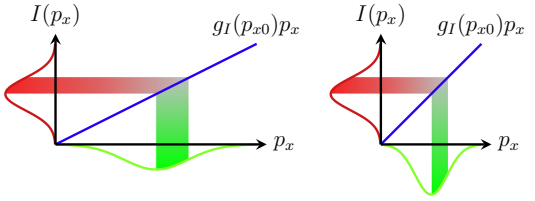


Fig. 6. Camera sensor noise n_I can be considered as additive white Gaussian noise (AWGN) which disturbs the intensity values $I(\boldsymbol{x}_R)$ and thus affects the disparity observation as AWGN. As shown on the left, for a low image gradient along the epipolar line, the influence of the sensor noise n_I is stronger than for a high image gradient.

intensity error. Therefore, the focus uncertainty σ_f^2 is defined as follows:

$$\sigma_f^2 = \alpha\cdot\frac{e_{ISS}(\hat{p}_x)}{g_I(\boldsymbol{x}_R(\hat{p}_x))^2} \tag{27}$$

The constant scaling factor α defines the weight of σ_f^2 with respect to $\sigma_{p_x}^2$. We chose α such that for micro lenses with a different focal length σ_f^2 equals on average to $\sigma_{p_x}^2$.

We define the overall observation uncertainty of the inverse virtual depth σ_z^2 as the sum of $\sigma_{p_x}^2$ and σ_f^2. From Eq. (15) one can see that z is the disparity p_x scaled by d^{-1}. Thus, for σ_z^2 the scaling factor d^{-2} has to be introduced, as given in Eq. (28).

$$\sigma_z^2 = d^{-2}\cdot\left(\sigma_{p_x}^2 + \sigma_f^2\right) \tag{28}$$

4.4. Updating inverse virtual depth hypothesis

As described in Section 4.2 the observations for the inverse virtual depth z are performed starting from the shortest baseline up to the largest possible baseline, for which a virtual image point is still seen in both micro images. In that way for each pixel an exhaustive stereo matching over all possible micro images is performed, leading to multi-view stereo. In our algorithm we incorporate new inverse virtual depth observations similar to the update step in a Kalman filter. Thus, the new inverse virtual depth distribution $\mathcal{N}(z,\sigma_z^2)$ results form the previous distribution $\mathcal{N}(z_p,\sigma_p^2)$ and the new inverse depth observation $\mathcal{N}(z_o,\sigma_o^2)$ as given in Eq. (29).

$$\mathcal{N}(z,\sigma_z^2) = \mathcal{N}\left(\frac{\sigma_p^2\cdot z_o + \sigma_o^2\cdot z_p}{\sigma_p^2+\sigma_o^2}, \frac{\sigma_p^2\cdot\sigma_o^2}{\sigma_p^2+\sigma_o^2}\right) \tag{29}$$

From Eq. (28) one can see that the inverse virtual depth variance σ_z^2 is proportional to d^{-2}. Furthermore, the number of observations increases with increasing baseline distance d. We consider M observations of a micro image point with disparity variance $\sigma_{p_x}^{2(i)} = \sigma_{p_x}^2$ $(i\in 1,2,\ldots,M)$ at the same baseline distance d. Therefore, the inverse virtual depth variance σ_z^2 results as given in Eq. (30).

$$\frac{1}{\sigma_z^2} = \sum_{i=1}^{M}\frac{d^2}{\sigma_{p_x}^{2(i)}} = \frac{M\cdot d^2}{\sigma_{p_x}^2} \tag{30}$$

Assuming that M and d are proportional to v one can see from Eq. (30) that σ_z^2 decreases with v^3.

4.5. Calculating a virtual depth map

Based on the observed inverse virtual depth z, a pixel in the raw image, defined by the coordinates $\boldsymbol{x}_R$, can be projected in a 3D space. We will call this 3D space the virtual image space and denote it by the coordinates $\boldsymbol{x}_V = (x_V, y_V, v = z^{-1})^T$. This transform is defined by an inverse central projection as follows:

$$x_V = (x_R - h_x)z^{-1} + h_x \tag{31}$$

$$y_V = (y_R - h_y)z^{-1} + h_y \tag{32}$$

Here $\boldsymbol{h} = (h_x, h_y)^T$ defines the center of the micro lens which projects the virtual image point $\boldsymbol{x}_V$ on the sensor. This inverse central projection is also visualized in Fig. 7. Defining the projection in homogeneous coordinates results in the following system of equations:

$$\begin{pmatrix} z \cdot x_V \\ z \cdot y_V \\ z \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & h_x & -h_x \\ 0 & 1 & h_y & -h_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_R \\ y_R \\ z \\ 1 \end{pmatrix} \tag{33}$$

Pixels which are projected to the same virtual image point are used to update the virtual depth of the virtual image point with the same probabilistic update method as described in Section 4.4.

5. Calibration of the plenoptic camera

To define the relation between virtual image space and object space a camera calibration has to be performed. This section presents several methods which we have developed to calibrate a focused plenoptic camera. The calibration is divided into two parts. In Section 5.1 we briefly present the method to calibrate the optical path. It is presented here for completeness, but it is not an original contribution of this paper. In Section 5.2 we present as original contribution the calibration of the depth map. Here we define the relationship between the object distance a_L and the virtual depth v.

5.1. Calibration of the optical path

The optical path defines the relationship between camera coordinates $\boldsymbol{x}_C = (x_C, y_C, z_C = a_L)^T$ and virtual image coordinates $\boldsymbol{x}_V = (x_V, y_V)$. In our approach we consider this projection as a central perspective projection, as it is for a regular camera. Thus, similar to a regular camera, our model is defined by a pinhole camera including a distortion model. The pinhole model is described by the equation:

$$\begin{pmatrix} \lambda \cdot x'_V \\ \lambda \cdot y'_V \\ \lambda \end{pmatrix} = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_C \\ y_C \\ z_C \end{pmatrix} \tag{34}$$

In Eq. (34) $\boldsymbol{x}'_V = (x'_V, y'_V)^T$ are the undistorted virtual image coordinates. Besides, f defines the principal distance (or camera constant) of the underlying pinhole camera model, while $\boldsymbol{c} = (c_x, c_y)^T$ defines the principal point. All three parameters, f, c_x, and c_y, are defined in pixels. We define the relation between the distorted points $\boldsymbol{x}_V$ and the undistorted points $\boldsymbol{x}'_V$ based on the commonly used model as presented by Brown (1966). The model consists of radial symmetric as well as radial asymmetric correction terms.

$$x_V = x'_V + \Delta x_{rad} + \Delta x_{tan} \tag{35}$$
$$y_V = y'_V + \Delta y_{rad} + \Delta y_{tan} \tag{36}$$

The radial symmetric correction terms Δx_{rad} and Δy_{rad} are defined as follows:

$$\Delta x_{rad} = (x'_V - c_x)\left(k_0 r^2 + k_1 r^4 + k_2 r^6\right) \tag{37}$$
$$\Delta y_{rad} = (y'_V - c_y)\left(k_0 r^2 + k_1 r^4 + k_2 r^6\right) \tag{38}$$
$$r = \sqrt{(x'_V - c_x)^2 + (y'_V - c_y)^2} \tag{39}$$

The radial asymmetric correction terms Δx_{tan} and Δy_{tan} are given by the following two equations:

$$\Delta x_{tan} = p_0\left(r^2 + 2(x'_V - c_x)^2\right) + 2p_1(x'_V - c_x)(y'_V - c_y) \tag{40}$$
$$\Delta y_{tan} = p_1\left(r^2 + 2(y'_V - c_y)^2\right) + 2p_0(x'_V - c_x)(y'_V - c_y) \tag{41}$$

We estimate the defined model based on a traditional, bundle adjustment based calibration method, as will be presented in the evaluation in Section 7.2. Of course the model could be further extended by introducing additional correction terms. Nevertheless we received good results with the presented model as will be shown later.

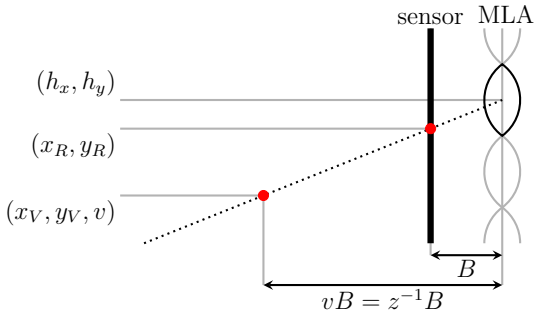


Fig. 7. Projection of a raw image point $\boldsymbol{x}_R$ to a virtual image point $\boldsymbol{x}_V$ based on an inverse central projection.

5.2. Calibration of the virtual depth

Purpose of the depth map calibration is to define the relationship between the virtual depth v supplied by the focused plenoptic camera and the metric object distance a_L.

As described in Section 2 the relationship between the virtual depth v and the object distance a_L relies on the thin lens equation, which is given in Eq. (1).

From Fig. 3 one can see, that the image distance b_L is linearly dependent on the virtual depth v. This dependency is as given by Eq. (42).

$$b_L = v \cdot B + b_{L0} \tag{42}$$

Substituting this in the thin lens equation (Eq. (1)) and rearranging the terms yields for $a_L(v)$:

$$a_L(v) = \left(\frac{1}{f_L} - \frac{1}{v \cdot B + b_{L0}}\right)^{-1} \tag{43}$$

This function depends on three unknown but constant parameters (f_L, B, and b_{L0}) which have to be estimated. This can be performed from a bunch of measured calibration points for which the object distance a_L is known. In this paper we present two novel model based calibration methods. For comparison the function will also be approximated by a curve fitting approach.

5.2.1. Method 1 – Physical model

The first model based approach estimates the unknown parameters of Eq. (43) explicitly. Since the main lens focal length is already known approximately from the lens specification, it is set as a constant value prior to the estimation. For each measured object distance $a_L^{\{i\}}$ the corresponding image distance $b_L^{\{i\}}$ is calculated based on the thin lens equation Eq. (1).

Since the image distance b_L is linearly dependent on the virtual depth v, the calculated image distances $b_L^{\{i\}}$ and the corresponding virtual depths $v^{\{i\}}$ are used to estimate the parameters B and b_{L0}. Eqs. (44)–(46) show the least squares estimation of the parameters.

$$\begin{pmatrix} \hat{B} \\ \hat{b}_{L0} \end{pmatrix} = \left(\boldsymbol{X}_{Ph}^T \cdot \boldsymbol{X}_{Ph}\right)^{-1} \cdot \boldsymbol{X}_{Ph}^T \cdot \boldsymbol{y}_{Ph} \tag{44}$$

$$\boldsymbol{y}_{Ph} = \begin{pmatrix} b_L^{\{0\}} & b_L^{\{1\}} & b_L^{\{2\}} & \cdots & b_L^{\{N\}} \end{pmatrix}^T \tag{45}$$

$$\boldsymbol{X}_{Ph} = \begin{pmatrix} v^{\{0\}} & v^{\{1\}} & v^{\{2\}} & \cdots & v^{\{N\}} \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}^T \tag{46}$$

Instead of solving for the parameters B and b_{L0} to fulfill the least squares criteria, one could also think of considering the inverse depth variances σ_z^2 to solve for a minimum variance.

5.2.2. Method 2 – Behavioral model

The second model based approach relies also on the function defined in Eq. (43). However, this method does not estimate the physical parameters explicitly as done in method 1, but a function which behaves similar to the physical model.

Eq. (43) can be rearranged to the term given in Eq. (47). Since the virtual depth v and the object distance a_L both are observable dimensions, a third variable $u = a_L \cdot v$ can be defined.

$$a_L = a_L \cdot v \cdot \frac{B}{f_L - b_{L0}} + v \cdot \frac{B \cdot f_L}{b_{L0} - f_L} + \frac{b_{L0} \cdot f_L}{b_{L0} - f_L} \tag{47}$$

Thus, from Eq. (47) the term given in Eq. (48) results. Here, the object distance a_L is defined as a linear combination of the measurable variables u and v.

$$a_L = u \cdot c_0 + v \cdot c_1 + c_2 \tag{48}$$

The coefficients c_0, c_1, and c_2 are defined as given in Eqs. (49)–(51).

$$c_0 = \frac{B}{f_L - b_{L0}} \tag{49}$$

$$c_1 = \frac{B \cdot f_L}{b_{L0} - f_L} \tag{50}$$

$$c_2 = \frac{b_{L0} \cdot f_L}{b_{L0} - f_L} \tag{51}$$

Since for Eq. (48) all three variables a_L, v, and u are observable dimensions, the coefficients c_0, c_1, and c_2 can be estimated based on a number of calibration points. For the experiments presented in Section 7.2 the coefficients c_0, c_1, and c_2 are estimated by using the least squares method as given in Eqs. (52)–(54).

$$\begin{pmatrix} \hat{c}_0 \\ \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \left(\boldsymbol{X}_{Be}^T \cdot \boldsymbol{X}_{Be} \right)^{-1} \cdot \boldsymbol{X}_{Be}^T \cdot \boldsymbol{y}_{Be} \tag{52}$$

$$\boldsymbol{y}_{Be} = \begin{pmatrix} a_L^{\{0\}} & a_L^{\{1\}} & \cdots & a_L^{\{N\}} \end{pmatrix}^T \tag{53}$$

$$\boldsymbol{X}_{Be} = \begin{pmatrix} a_L^{\{0\}} v^{\{0\}} & a_L^{\{1\}} v^{\{1\}} & \cdots & a_L^{\{N\}} v^{\{N\}} \\ v^{\{0\}} & v^{\{1\}} & \cdots & v^{\{N\}} \\ 1 & 1 & \cdots & 1 \end{pmatrix}^T \tag{54}$$

After rearranging Eq. (48) the object distance a_L can be described as a function of the virtual depth v and the estimated parameters c_0, c_1, and c_2 as given in Eq. (55).

$$a_L(v) = \frac{v \cdot c_1 + c_2}{1 - v \cdot c_0} \tag{55}$$

5.2.3. Method 3 – Curve fitting

The third method is presented for comparison purpose and is a common curve fitting approach. It approximates the function between the virtual depth v and the object distance a_L without paying attention to the function defined in Eq. (43).

It is known that any differentiable function can be represented by a Taylor-series and thus, by a polynomial of infinite order. Hence, in the approach presented here the functions which describes the object distance a_L depending on the virtual depth v will be defined as a polynomial as well. A general definition of this polynomial is given in Eq. (56).

$$a_L(v) \approx \sum_{k=0}^{K} l_k \cdot (v)^k \tag{56}$$

Similar to the second method the polynomial coefficients l_0 to l_K are estimated based on a bunch of calibration points. In the experiments presented in Section 7 a least squares estimator as given in Eqs. (57)–(59) is used.

$$\begin{pmatrix} \hat{l}_0 \\ \hat{l}_1 \\ \vdots \\ \hat{l}_K \end{pmatrix} = \left(\boldsymbol{X}_{Pol}^T \cdot \boldsymbol{X}_{Pol} \right)^{-1} \cdot \boldsymbol{X}_{Pol}^T \cdot \boldsymbol{y}_{Pol} \tag{57}$$

$$\boldsymbol{y}_{Pol} = \begin{pmatrix} a_L^{\{0\}} & a_L^{\{1\}} & \cdots & a_L^{\{N\}} \end{pmatrix}^T \tag{58}$$

$$\boldsymbol{X}_{Pol} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ v^{\{0\}} & v^{\{1\}} & \cdots & v^{\{N\}} \\ v^{\{0\}2} & v^{\{1\}2} & \cdots & v^{\{N\}2} \\ \vdots & \vdots & \ddots & \vdots \\ v^{\{0\}M} & v^{\{1\}M} & \cdots & v^{\{N\}M} \end{pmatrix}^T \tag{59}$$

For this method a trade-off between the accuracy of the approximated function and the order of the polynomial has to be found. A high order of the polynomial results in more effort for calculating an object distance from the virtual depth. Besides, for high orders the matrix inversion as defined in Eq. (57) results in numerical inaccuracy. For such cases a different method for solving the least squares problem has to be used (e.g. Cholesky decomposition).

For all depth calibration methods we do not define a distortion model for the virtual depth, as it has been done for instance by Johannsen et al. (2013). As the results in Section 7.2.3 will show, there occurs almost no image distortion. Thus, we also neglect the virtual depth distortion compared to the stochastic depth error.

6. Plenoptic camera based visual odometry

The plenoptic camera based visual odometry that is presented here is a direct and semi-dense method. Direct means it does not perform any feature extraction, but works directly on pixel intensities. The algorithm is called semi-dense since it works only on pixels with sufficient gradient and neglects homogeneous regions in the intensity image. The method is based on the monocular approach published by Engel et al. (2013, 2014). It has been adapted and modified by us to be suitable for plenoptic cameras.

The method presented in this section uses both, the probabilistic virtual depth map, which is established as described in Section 4 and the synthesized totally focused intensity image (see Perwaß and Wietzke, 2012). In addition, the virtual depth calibration is used to receive a metric scale of the scene. For the complete section we will always consider the images to be undistorted. Thus, a point is always defined by its undistorted virtual image coordinates $\boldsymbol{x}'_V = (x'_V, y'_V)^T$.

This part about the plenoptic camera based visual odometry is structured as follows: In Section 6.1 we introduce a probabilistic metric depth model which is defined similar to the virtual depth model in Section 4. Section 6.2 briefly describes the initialization of the method. While Sections 6.3 and 6.4 present the tracking of new frames and updating of the depth map, Section 6.5 describes how the existing metric depth map is propagated from reference frame to reference frame.

A complete work-flow of the plenoptic camera based visual odometry is shown in Algorithm 1.

Algorithm 1. Plenoptic camera based visual odometry

Initialization:
- define first recorded frame as reference frame
- estimate probabilistic inverse virtual depth map $Z_{Ref}(\boldsymbol{x}'_V)$
- synthesize totally focused image of reference frame $I_{Ref}(\boldsymbol{x}'_V)$
- initialize probabilistic inverse metric depth map $D(\boldsymbol{x}'_V)$ based on $Z_{Ref}(\boldsymbol{x}'_V)$

while new frame is present **do**

Track new frame:
- estimate probabilistic inverse virtual depth map $Z_{New}(\boldsymbol{x}'_V)$ of new frame
- synthesize totally focused image of new frame $I_{New}(\boldsymbol{x}'_V)$
- estimate rigid body transform $G(\xi)$ from reference frame to new frame, based on $I_{Ref}(\boldsymbol{x}'_V), I_{New}(\boldsymbol{x}'_V)$, and $D(\boldsymbol{x}'_V)$

Update inverse metric depth:
- perform stereo matching from $I_{Ref}(\boldsymbol{x}'_V)$ to $I_{New}(\boldsymbol{x}'_V)$ for each valid depth pixel in $D(\boldsymbol{x}'_V)$
  → results in new depth observations $d_o(\boldsymbol{x}'_V)$
- calculate inverse depth variance $\sigma_o^2(\boldsymbol{x}'_V)$ for each new observation $d_o(\boldsymbol{x}'_V)$
- incorporate new observations ($d_o(\boldsymbol{x}'_V)$ and $\sigma_o^2(\boldsymbol{x}'_V)$) in probabilistic inverse metric depth map $D(\boldsymbol{x}'_V)$

if new reference frame is needed **then**

Set new reference frame:
- set last tracked frame as new reference frame
  - $I_{Ref}(\boldsymbol{x}'_V) \longleftarrow I_{New}(\boldsymbol{x}'_V)$
  - $Z_{Ref}(\boldsymbol{x}'_V) \longleftarrow Z_{New}(\boldsymbol{x}'_V)$
- propagate inverse metric depth $D(\boldsymbol{x}'_V)$ to new reference frame using $G(\xi)$ of last tracked frame
- initialize new pixels in $D(\boldsymbol{x}'_V)$
  - pixels which were not visible in last reference frame
  - based on inverse virtual depth $Z_{Ref}(\boldsymbol{x}'_V)$ of new reference frame

end

end

6.1. Inverse metric depth map

Similar to the inverse virtual depth map an inverse metric depth map is defined. The inverse depth of each pixel is modeled as a Gaussian distributed random variable. This is done much the same as for the inverse virtual depth since for a pair of pinhole cameras the inverse metric depth is approximately proportional to the estimated disparity of corresponding points in the two frames, at least for a rotation matrix close to a unity matrix. For each pixel $\boldsymbol{x}'_V = (x'_V, y'_V)^T$ in the undistorted inverse metric depth map, the inverse depth value $d(\boldsymbol{x}'_V)$ is calculated from the virtual depth, as given in Eq. (60). Here c_0, c_1, and c_2 are the coefficients received from the camera calibration, as presented in Section 5.2.2.

$$d(\boldsymbol{x}'_V) = z_C^{-1}(\boldsymbol{x}'_V) = \frac{1 - v(\boldsymbol{x}'_V)\cdot c_0}{v(\boldsymbol{x}'_V)\cdot c_1 + c_2} = \frac{z(\boldsymbol{x}'_V) - c_0}{c_1 + c_2 \cdot z(\boldsymbol{x}'_V)} \tag{60}$$

In Eq. (60) $z(\boldsymbol{x}'_V) = v^{-1}(\boldsymbol{x}'_V)$ defines the inverse virtual depth of the pixel $\boldsymbol{x}'_V$. Of course any other calibration model could be used to calculate the inverse metric depth. The corresponding variance $\sigma_d^2(\boldsymbol{x}'_V)$ is received from the inverse virtual depth variance as follows:

$$\sigma_d^2(\boldsymbol{x}'_V) = \left|\frac{\partial d}{\partial z}\right|^2 \cdot \sigma_z^2(\boldsymbol{x}'_V) = \frac{(c_1 + c_0 \cdot c_2)^2}{(c_1 + z(\boldsymbol{x}'_V)\cdot c_2)^4} \cdot \sigma_z^2(\boldsymbol{x}'_V) \tag{61}$$

Due to the nonlinear projection from the inverse virtual depth $z(\boldsymbol{x}'_V)$ to the inverse metric depth $d(\boldsymbol{x}'_V)$, these calculated inverse depth values do not result from a Gaussian process anymore. Thereby we accept a small error when initializing the inverse metric depth map. Nevertheless, as will be described in the following these values are used only as initial depth hypotheses which will be continuously updated by new stereo observations in the totally focused images. These new observations result from a Gaussian process and in general are much more reliable than the initial values. Thereby the initial error becomes neglectable.

6.2. Initialization

When running the algorithm the first recorded frame is set as reference frame. For this frame the probabilistic inverse virtual depth map is estimated as presented in Section 4 and the totally focused image is synthesized. In the following step, for the reference frame the probabilistic inverse metric depth map ($d(\boldsymbol{x}'_V)$ and $\sigma_d^2(\boldsymbol{x}'_V)$) of the reference frame is initialized based on the probabilistic inverse virtual depth, as described in Section 6.1. Here, we initialize only the pixels which have a valid inverse virtual depth estimate. These are all the pixels which also have sufficient intensity gradient.

Thereby, our first reference frame is defined by its totally focused intensity image $I_{Ref}(\boldsymbol{x}'_V)$ as well as its probabilistic inverse metric depth map $D(\boldsymbol{x}'_V)$.

$$D(\boldsymbol{x}'_V) \sim \mathcal{N}\left(d(\boldsymbol{x}'_V), \sigma_d^2(\boldsymbol{x}'_V)\right) \tag{62}$$

6.3. Tracking of new frames

For each newly recorded frame again the totally focused intensity image $I_{New}(\boldsymbol{x}'_V)$ is synthesized. Since for the image synthesis the virtual depth map is needed, we estimate the complete inverse virtual depth map for each recorded frame and synthesize the image based on that depth map. However, in the future it could be considered to use the already existing depth map of the reference frame and project it into the newly recorded frame. This works well if the frame rate is high with respect to the camera motion since the pose of the new frame (relative to the reference frame) has to be known, at least roughly.

Based on the synthesized image the frame pose with respect to the current reference frame is estimated. The transform between the camera coordinates of the reference frame $\boldsymbol{x}_W$ and the camera coordinates of the new frame $\boldsymbol{x}_C$ is defined by a rigid body transform $\boldsymbol{G} \in SE(3)$, as given in Eq. (63).

$$\boldsymbol{x}_C = \begin{pmatrix} x_C \\ y_C \\ z_C \\ 1 \end{pmatrix} = \boldsymbol{G}\cdot\boldsymbol{x}_W = \boldsymbol{G}\cdot\begin{pmatrix} x_W \\ y_W \\ z_W \\ 1 \end{pmatrix} \tag{63}$$

The rigid body transform $\boldsymbol{G}$ is defined as the combination of a rotation and a translation in 3D space:

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix} \in SE(3) \tag{64}$$

The Matrix $\boldsymbol{G}$ has six degrees of freedom. These degrees of freedom are in the following combined in the six-dimensional vector $\xi \in \mathbb{R}^6$, which has to be estimated.

The vector ξ is optimized by minimizing the sum of squared photometric errors $r_i(\xi)^2$, over all valid depth pixels, between the synthesized image of the new frame $I_{New}(\boldsymbol{x}'_V)$ and the one of the reference frame $I_{Ref}(\boldsymbol{x}'_V)$. The squared photometric error, for a certain pixel $\boldsymbol{x}'^{(i)}_V$ in the reference frame, can be defined as follows:

$$r_i(\xi)^2 := \left(I_{Ref}\left(\boldsymbol{x}_V'^{(i)}\right) - I_{New}\left(w\left(\boldsymbol{x}_V'^{(i)}, d(\boldsymbol{x}_V'^{(i)}), \xi\right)\right)\right)^2 \quad (65)$$

Here the warping of a pixel in the reference frame to its pixel coordinates in the new frame is combined in the function $w(\boldsymbol{x}_V', d(\boldsymbol{x}_V'), \xi)$. This function is defined by the projection model of the camera.

In this nonlinear optimization process, which in our case is solved by the Levenberg–Marquardt algorithm, a weighting scheme is used that handles the different inverse depth variances of the pixels in the probabilistic depth map as well as outliers which result for instance from occlusion. The robustness of the method is improved by performing the optimization on different pyramid levels, starting from a low image resolution up to the full image resolution.

For a further explanation we refer to Engel et al. (2014) where this optimization process is described in detail.

6.4. Updating the inverse metric depth

The probabilistic inverse metric depth map $D(\boldsymbol{x}_V')$ of the current reference frame is updated based on each newly tracked frame. Therefore, for each pixel $\boldsymbol{x}_V'$ with a valid inverse metric depth hypothesis, stereo-matching along the epipolar line is performed. Here, stereo matching is performed by minimizing the sum of squared intensity differences of a one-dimensional pixel patch along the epipolar line. Sub-pixel accuracy is achieved by linear interpolation between the samples of the totally focused image.

Similar to the virtual depth estimation (Section 4.3.1) the depth range is limited to $d(\boldsymbol{x}_V') \pm n\sigma_d(\boldsymbol{x}_V')$.

Engel et al. (2013) show how the error of an observed disparity can be modeled by two different error sources. One error source is the geometric error, which results from noise on the estimated camera pose and on the intrinsic camera parameters. It affects the position and orientation of the epipolar line. The second error source is the photometric error, which results from noise in the intensity image. It is considered that both errors are Gaussian distributed and additively interfere the observed disparity and thus the observed inverse metric depth. Thus, based on these two error sources, beside the inverse metric depth observation $d_o(\boldsymbol{x}_V')$ itself, a corresponding quality criteria, the inverse depth variance $\sigma_o^2(\boldsymbol{x}_V')$, can be defined.

For each valid depth pixel $\boldsymbol{x}_V'$ the new observation of the inverse metric depth is incorporated into the already existing inverse metric depth hypothesis $D(\boldsymbol{x}_V')$. Therefore, again the Kalman filter step, as already given in Eq. (29), is applied.

6.5. Propagating the inverse metric depth

With changing perspective of the camera, the number of valid depth pixels $\boldsymbol{x}_V'$ in the reference frame, which can be mapped to the corresponding pixel in the newly tracked frame, decreases. Therefore, at one point a new reference frame has to be selected. Thus, the last tracked frame will be set as the new reference frame.

The decision for changing the reference frame is made based on a score which considers the length of the translation vector $\boldsymbol{t}$ between the current reference and the new frame as well as the percentage of valid depth pixels used for tracking.

After changing the reference frame the algorithm still is supposed to benefit from the continuously updated depth map of the old reference frame. Therefore, the probabilistic depth map of the old reference frame $D_{Old}(\boldsymbol{x}_V')$ is propagated to the new frame.

Here, the pixel coordinates in the new frame can be calculated based on the mapping function $w(\boldsymbol{x}_V', d(\boldsymbol{x}_V'), \xi)$:

$$\boldsymbol{x}_{VNew}'^{(i)} = w\left(\boldsymbol{x}_V'^{(i)}, d^{(i)}, \xi\right) \quad (66)$$

Besides, the new inverse metric depth $d_{New}(\boldsymbol{x}_V')$ is just the inverse of the component z_C of the respective camera coordinates.

Under the assumption that the rotation between the old and new reference frame is small, the following approximation holds:

$$d_{New}(\boldsymbol{x}_V') \approx \left(\frac{1}{d_{Old}(\boldsymbol{x}_V')} + t_z\right)^{-1} \quad (67)$$

Considering the approximation given in Eq. (67), based on the propagation of uncertainty, the new inverse metric depth variance $\sigma_{dNew}^2(\boldsymbol{x}_V')$ can be calculated as follows:

$$\sigma_{dNew}^2(\boldsymbol{x}_V') \approx \left(\frac{d_{New}(\boldsymbol{x}_V')}{d_{Old}(\boldsymbol{x}_V')}\right)^4 \cdot \sigma_{dOld}^2(\boldsymbol{x}_V') \quad (68)$$

After propagating the depth to the new reference frame, all pixels which have a valid inverse virtual depth value but where not seen in the old reference frame are initialized as described in Section 6.2. Subsequently the algorithm continues tracking new frames and updating the inverse depth map of the reference frame.

7. Evaluation of the proposed methods

This section presents the evaluation of all proposed methods. Here, we first evaluate the virtual depth estimation algorithm presented in Section 4 itself and compare it to a conventional algorithm, as shown in Section 7.1.

In Section 7.2 the evaluation of the camera calibration as described in Section 5 is presented. Here we also evaluate the accuracy of the metric depth which is received from the estimated virtual depth.

In Section 7.3 we put everything together for the focused plenoptic camera based visual odometry. We present the 3D point cloud of a sample scene which was generated by this visual odometry approach. Additionally, we show how the depth accuracy compared to that of a single light-field frame can be improved.

All experiments which are presented in the following sections are performed based on a Raytrix R5 camera with a main lens focal length of $f_L = 35$ mm. The camera has a sensor resolution of 2048×2048 pixel at a pixel pitch of 5.5 mm. Thus, the camera has a FOV of approximately 18°.

7.1. Evaluation of the virtual depth estimation

This section presents the evaluation of our proposed depth estimation algorithm which in the following will be abbreviated by MVS (multi-view stereo). For comparison we perform all experiments for our MVS and a conventional block matching algorithm (BMA) as it is used by Perwaß and Wietzke (2012).

7.1.1. Experiments

To evaluate the depth estimation methods a planar target is recorded for different object distances, as shown in Fig. 8. Here, the images show the view of the camera for two different target distances. Since the target is placed frontal to the plenoptic camera for a perfect estimation one would expect a constant virtual depth across the complete plane.

For each of the recorded frames a virtual depth map is calculated, by using both our probabilistic MVS method and the conventional BMA. Since we only want to evaluate the depth estimation algorithm itself, no post processing steps like filtering or hole filling have been performed on the virtual depth maps.

Since our method offers additionally to the virtual depth $v = z^{-1}$ an inverse virtual depth variance σ_z^2, two different depth maps were calculated based on our MVS algorithm. While the first depth map considers all valid depth pixel disregarding their variance, the second depth map considers only those depth pixel

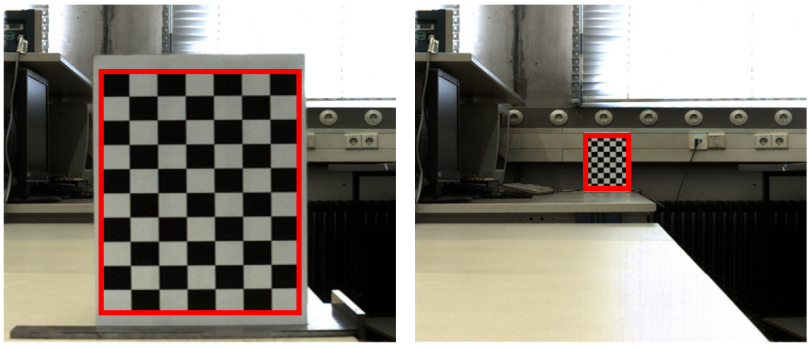


Fig. 8. Setup to evaluate the virtual depth estimation. The virtual depth estimation algorithms (MVS and BMA) are evaluated based on a planar chessboard target for different object distances. Only depth values in the red marked region of interest are evaluated.

which have a variance σ_z^2 underneath a certain threshold $T(z)$, as defined in Eq. (69).

$$\sigma_z^2(\boldsymbol{x}_V) < T(z) = \beta \cdot z(\boldsymbol{x}_V)^3 \tag{69}$$

The threshold $T(z)$ is chosen as a third order function of z due to the thoughts made in Section 4.4. In Eq. (69) β is just a scaling factor, which defines the point density of the resulting depth map. In our experiments a scaling factor $\beta = 0.1$ was chosen. This resulted in a more or less equal point density for our approach compared to the BMA. It is important to emphasize, that here no low-pass filtering is performed and just uncertain estimates are removed.

The MVS performs matching based on a one-dimensional patch which is 5 pixels long. Since we are able to define a continuous cost function by linear interpolation between the samples, the disparity estimation of the MVS is not restricted to any regular sub-pixel grid.

The BMA is performed with three different settings. For all settings a block diameter of 4 pixel was selected. Besides, for the three settings disparities are estimated with sub-pixel accuracies of 0.1 pixel ($BMA_{(0.1)}$), 0.25 pixel ($BMA_{(0.25)}$), and 0.5 pixel ($BMA_{(0.5)}$).

7.1.2. Results

Fig. 9 exemplary shows the depth maps calculated for an object distance $a_L = 1.2$ m. These depth maps correspond to the scene which is shown on the left side in Fig. 8. Fig. 9(a) and (b) show the results of our MVS algorithm. Here, Fig. 9(a) includes all valid depth pixels, while Fig. 9(b) includes only those which have a variance $\sigma_z^2 < T(z)$, as defined in Eq. (69). The depth map in Fig. 9(c) shows the results of the conventional BMA with a sub-pixel accuracy of 0.25 pixel.

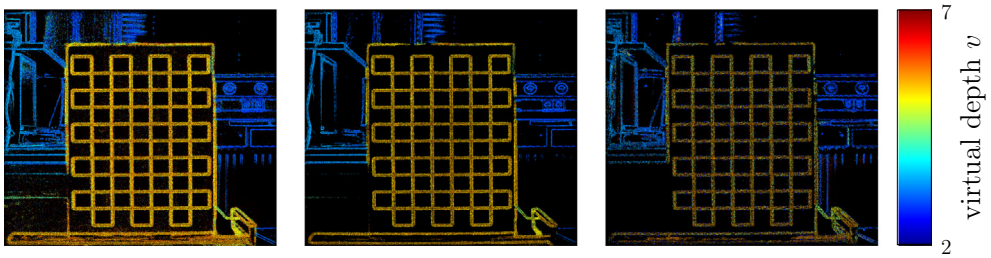
From Fig. 9 one can already see, that the outliers in our method are drastically reduced by introducing the threshold $T(z)$, while most of the details are kept. Besides, one can see that the depth map of the BMA is much sparser than the raw depth map resulting from the MVS. In addition it seems that the outliers of the BMA, especially on the chessboard plane are not statistically independent, but occur in clusters.

Beside the qualitative evaluation based on the color coded depth maps some statistics were calculated for different object distances a_L. In this Section we present the results for $a_{L1} = 1.2$ m, $a_{L2} = 3.1$ m, and $a_{L3} = 5.1$ m. For all three object distances Table 1 shows the depth pixel density of the corresponding algorithm. The depth pixel density is defined as the ratio between the number of valid depth pixels and the total number of pixels within the region of interest.

One can see, that our method has a higher depth pixel density than the BMA for all object distances.

For all three object distances we calculated the empirical standard deviation of the inverse virtual depth values $z = v^{-1}$ across the chessboard target. The results are shown in Table 2. As one can see, the standard deviation of our MVS approach is better than that of the BMA for all three object distances, even without removing outliers. After removing outliers, we achieve a standard deviation which is at least three times better than that of the BMA, while still having a higher depth pixel density (see Table 1). It is also quite interesting to see that only slightly reducing the depth pixel density, by introducing the threshold $T(z)$, highly reduces the empirical standard deviation of the inverse virtual depth.

Figs. 10 and 11 show exemplary the virtual depth histograms across the chessboard target, for the MVS and the BMA with a sub-pixel accuracy of 0.25 pixel, for the object distances $a_{L1} = 1.2$ m and $a_{L3} = 5.1$ m. Especially from Fig. 10 one can see that the outliers of the BMA have some systematic characteristic. The histograms again show quite well how the outliers in our approach are removed by introducing the threshold $T(z)$.



(a) MVS (all depths) (b) MVS ($\sigma_z^2 < T(z)$) (c) $BMA_{(0.25)}$

Fig. 9. Color coded virtual depth maps calculated from the raw image of a Raytrix R5 camera. (a) Depth map calculated based on our MVS algorithm. All valid depth pixels are considered. (b) Depth map calculated based on our MVS algorithm. Only depth pixels with a variance $\sigma_z^2 < T(z)$ are considered. (c) Depth map calculated by the conventional BMA with a sub-pixel accuracy of 0.25 pixel.

Table 1
Depth pixel density on the chessboard target for different object distances a_L.

Method	Depth pixel density		
	a_{L1}	a_{L2}	a_{L3}
MVS (all depths)	0.3077	0.5041	0.5647
MVS ($\sigma_z^2 < T(z)$)	0.1790	0.3900	0.4769
$BMA_{(0.1)}$	0.0609	0.0824	0.1039
$BMA_{(0.25)}$	0.1572	0.3018	0.4232
$BMA_{(0.5)}$	0.1383	0.2640	0.4768

Table 2
Empirical standard deviation of the inverse virtual depth z for different object distances a_L.

Method	Standard deviation		
	a_{L1}	a_{L2}	a_{L3}
MVS (all depths)	0.0366	0.0505	0.0385
MVS ($\sigma_z^2 < T(z)$)	0.0104	0.0167	0.0169
$BMA_{(0.1)}$	0.0840	0.0856	0.0975
$BMA_{(0.25)}$	0.0772	0.0644	0.0682
$BMA_{(0.5)}$	0.0645	0.0586	0.0530

7.2. Evaluation of the calibration methods

We show in Section 7.2.1 the setups we used for the calibration of the optical path as well as for the depth. In Section 7.2.2 we present how the parameters for the optical path model are obtained. Different experiments which compare the three depth calibration methods to each other are also presented here. The results of both, optical path and depth calibration are presented in Section 7.2.3.

7.2.1. Calibration setups

Fig. 12 shows the setups which were used to calibrate the plenoptic camera. For the calibration of the optical path a 3D calibration target was used, as shown in Fig. 12(a). This target consists of a number of coded and uncoded calibration points which can be detected automatically from the recorded images.

For the depth calibration we used the setup as shown in Fig. 12 (b). Here we have a chessboard calibration target which consists of 7×10 fields and thus results in 54 reference points (intersections between fields). This calibration target is moved along the optical axis of the camera. Behind the camera a laser rangefinder (LRF) is assembled which measures a metric reference distance to the target.

7.2.2. Experiments

7.2.2.1. Calibration of the optical path. The parameters of our optical path model as defined in Section 5.1 were estimated based on a professional calibration software. This software automatically detects the circular target points in the recorded images and performs a bundle adjustment. For the calibration performed here,

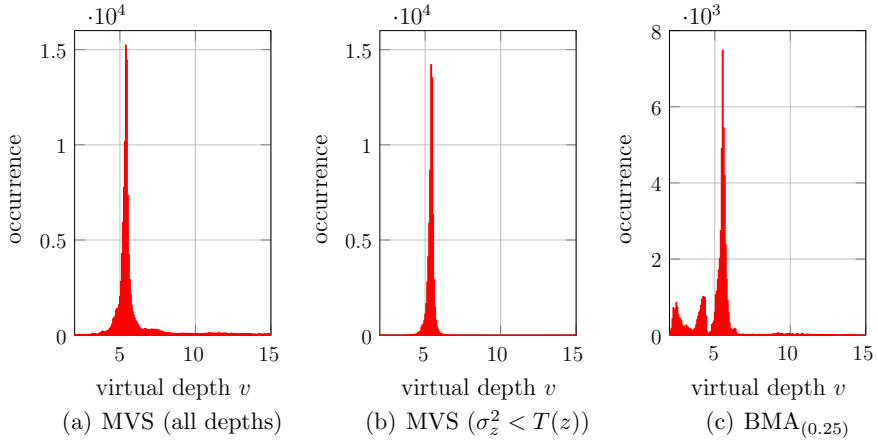


Fig. 10. Virtual depth histograms for object distance $a_{L1} \approx 1.2$ m. (a) Histogram of our MVS algorithm including all valid depth pixels. (b) Histogram of our MVS algorithm including all depth pixels with $\sigma_z^2 < T(z)$. (c) Histogram of the conventional BMA with a sub-pixel accuracy of 0.25 pixel.

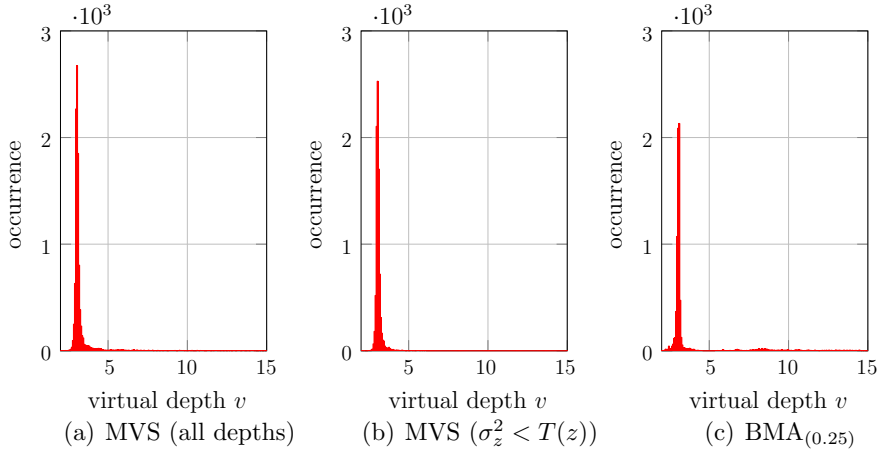


Fig. 11. Virtual depth histograms for object distance $a_{L1} \approx 5.1$ m. (a) Histogram of our MVS algorithm including all valid depth pixels. (b) Histogram of our MVS algorithm including all depth pixels with $\sigma_z^2 < T(z)$. (c) Histogram of the conventional BMA with a sub-pixel accuracy of 0.25 pixel.

(a) Optical path calibration setup (b) Depth calibration setup

Fig. 12. Setups to calibrate the focused plenoptic camera and to measure depth accuracy. (a) Setup for the calibration of the optical path. (b) Setup for the calibration of the depth map.

91 images were recorded from as different as possible perspectives.

7.2.2.2. Calibration of the depth map. To evaluate the different depth calibration methods, based on the setup shown in Fig. 12(b) a series of measurements was recorded. Therefor the chessboard target was recorded for different object distances.

For this series one cannot guarantee, that the target is always parallel to the sensor plane. Thus, the measured distance does not hold true for all points on the plane. Nevertheless, since we are able to detect the chessboard corner points in the image and we know their position on the target except for some scaling factor s, we can define the relation between coordinates on the chessboard $\boldsymbol{x}_{CB} = (x_{CB}, y_{CB}, z_{CB})^T$ and the undistorted virtual image coordinates $\boldsymbol{x}'_V = (x'_V, y'_V)^T$:

$$\begin{pmatrix} \lambda \cdot x'_V \\ \lambda \cdot y'_V \\ \lambda \end{pmatrix} = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix} \cdot (s\boldsymbol{R} \quad \boldsymbol{t}) \cdot \begin{pmatrix} x_{CB} \\ y_{CB} \\ z_{CB} \\ 1 \end{pmatrix} \tag{70}$$

Here, $\boldsymbol{R}$ defines the 3D rotation and $\boldsymbol{t}$ the 3D translation from chessboard coordinates to camera coordinates. Since we know the intrinsic camera parameters (f, c_x, and c_y), based on the undistorted recorded image point $\boldsymbol{x}'_V$, the rotation matrix $\boldsymbol{R}$ and the translation vector with respect to the scaling factor $\boldsymbol{t}' = \boldsymbol{t} \cdot s^{-1}$ can be estimated. For the case that the chessboard coordinates $\boldsymbol{x}_{CB}$ are defined with the x-y-plane being on the chessboard plane ($z_{CB} = 0$), the third coefficient of the translation vector equals the object distance $\boldsymbol{t}(3) = a_L$. Between the scaled object distance $a'_L = a_L \cdot s^{-1}$ and the target distances measured by the LRF d_{LRF} the following linear relation can be defined:

$$d_{LRF} = s \cdot a'_L + a_{L0} \tag{71}$$

Here, both, the scaling factor s and the zero point offset (between LRF and camera) are constant and thus can be estimated based on all measured target distances. In that way for each point on the chessboard plane a very precise object distance a_L can be calculated.

The chessboard pattern was recorded at 48 different object distances a_L in the range from 0.85 m to 5.02 m. Here all object distances are more or less uniformly distributed over the complete range, while the spacing between two distances ranges from 5 cm to 10 cm. Since the pattern on the calibration target has 54 reference points, 54 measurement points are received for each recorded target. For each measurement point the object distance a_L is calculated separately as described above.

Beside the physical model (Section 5.2.1) and the behavioral model (Section 5.2.2) based calibration method, the curve fitting approach (Section 5.2.3) was performed by using a third and a sixth order polynomial.

In a first experiment only the measured object distances of five recorded targets were used for calibration. This experiment was performed to evaluate if a low number of calibration points is sufficient to receive reliable calibration results.

In a second experiment only the measured points with an object distance of less than 2.9 m were used for calibration. In this experiment it was supposed to investigate how strong the estimated functions are drifting off from the measured data outside the range of calibration.

To evaluate the accuracy of the depth a third experiment was performed. Based on all measured points the root mean square error (RMSE) with respect to the distance of the calibration target was calculated. The object distances, which were calculated from the virtual depth, were converted to metric object distance by using the behavioral model presented in Section 5.2.2.

In the experiments we use the depth map calculated based on our virtual depth estimation method. The only post processing we do is filtering of the inverse virtual depth z with a weighted average filter, as shown in Eq. (72).

$$z_{AV} = \frac{\sum_i z_i \cdot \left(\sigma_{zi}^2\right)^{-1}}{\sum_i \left(\sigma_{zi}^2\right)^{-1}} \tag{72}$$

Here the weights are defined by the inverse of the variance σ_z^2 of a pixel.

In this experiment the parameters of the behavioral model were estimated based on all measured object distances, which then also were used for evaluation.

7.2.3. Results

7.2.3.1. Calibration of the optical path. The calibration of the optical path resulted in the intrinsic parameters as given in Table 3.

These parameters conform to what we expected. The principal point was expected to be somewhere around the center of the image. Since the virtual images $I(\boldsymbol{x}_V)$ used for calibration have a size of 1024×1024 pixel, this conforms quite well. The principal distance of the pinhole camera model f should be within the same order of magnitude as the main lens focal length f_L. For the pixels of the virtual image we assume a size 11×11 µm (double the width and height of the pixels in the raw image). This assumption holds since the virtual image has half as many pixels as the raw image in width and height respectively. Thus, a pinhole camera principal distance (in millimeter) $f_{PH} = 34,83$ mm is received. This conforms quite well to $f_L = 35$ mm.

For the distortion model the parameters given in Tables 4 and 5 are received.

From those parameters one can see, that for the 35 mm focal length, there occurs almost no distortion. The calibration resulted

Table 3
Intrinsic camera parameters.

f (pixel)	c_x (pixel)	c_y (pixel)
3166.44	510.98	516.31

Table 4
Radial symmetric distortion parameters.

k_0 (pixel^{-2})	k_1 (pixel^{-4})	k_2 (pixel^{-6})
-2.13×10^{-8}	-1.13×10^{-14}	-2.99×10^{-21}

Table 5
Radial asymmetric distortion parameters.

p_0 (pixel^{-1})	p_1 (pixel^{-1})
-1.60×10^{-10}	-3.07×10^{-11}

in a projection error with RMSE of 0.0438 pixel in x- and 0.0494 pixel in y-direction.

7.2.3.2. Calibration of the depth map. Fig. 13 shows the results corresponding to the first experiment. The red dots represent the calibration points for the five object distances. The green dots are the remaining measured points which were not used for calibration. As one can see, the physical model as well as the behavioral model are almost congruent. Both curves match the measured distances very well over the whole range from 0.85 m to 5.02 m. For the polynomials of order three and six instead, five object distances are not sufficient to approximate the function between virtual depth v and object distance a_L accurately. Both functions fit to the points used for calibration but do not define the underlying model properly in between the calibration points.

Fig. 14 shows the results of the second experiment. Again, the points used for calibration are represented as red dots and the green dots represent the remaining points. Both model based calibration approaches are again almost congruent and describe very well the measured distances in the range of calibration up to 2.9 m. In this range also the estimated polynomials fit the measured distances very well. Nevertheless, for object distances larger than 2.9 m especially the third order but also the sixth order polynomial are drifting away from the measurements. The functions of both model based approaches still match the measured values very well up to an object distance of 5.02 m. Though, the physical model still fits the data a little bit better than the behavioral model which has one more degree of freedom. The results show, that both model based functions are able to convert the virtual depth to an object distance even outside the range of calibration points with good accuracy.

As mentioned in Section 7.2.2.2, to evaluate the depth resolution of the plenoptic camera the complete series of measurement points was used for both, the prior calibration as well as for the evaluation. For each of the 48 object distances the RMSE is calculated. The RMSE is calculated based on all 54 measurement points per object distance. Therefor the error is defined as the difference between the measured object distance a_L^{Ref} and the one calculated based on the virtual depth $a_L(v)$.

Fig. 15 shows the RMSE of the metric depth as function of the object distance a_L. Besides, the figure shows the simulated accuracy for a similar camera setup as the one used in the experiments. In the simulation a maximum focus distance of 10 meters was set. For the estimated disparity a standard deviation of $\sigma_{p_x} = 0.1$ pixel was selected. This seems to be a plausible value, since one depth estimate results as the combination of multiple observations. From the graph one can see that the measured results conform comparatively well with the simulation, even though a very low σ_{p_x} was chosen.

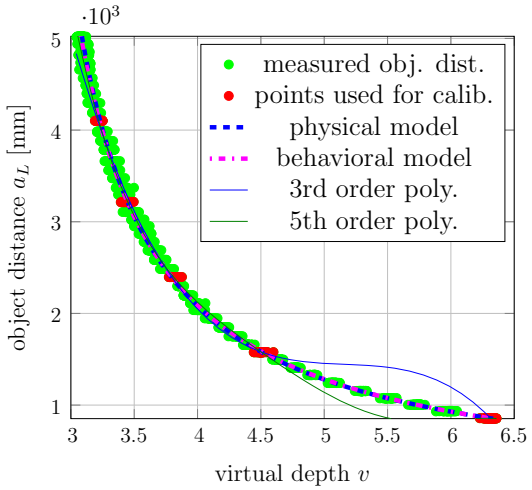


Fig. 13. Results of the depth calibration using only five object distances. The calibration was performed based on the physical model, the behavioral model and the curve fitting approach.

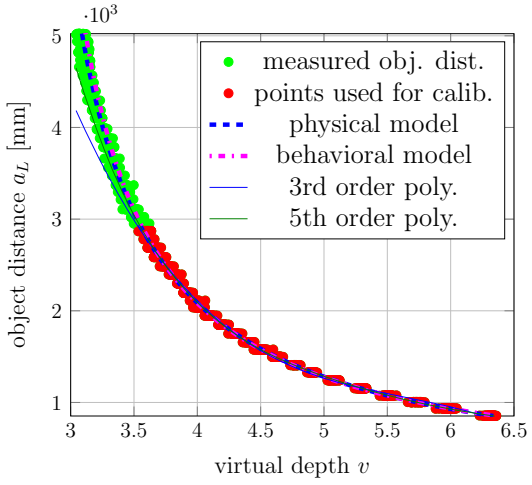


Fig. 14. Results of the depth calibration using object distances up to 2.9 m. The calibration was performed based on the physical model, the behavioral model and the curve fitting approach.

7.3. Evaluation of the plenoptic camera based visual odometry

This section presents the evaluation of our proposed visual odometry algorithm. Here we incorporate both, the virtual depths estimated based on the method presented in Section 4 and the camera model received from the calibration, as presented in Section 5.

7.3.1. Experiments

To evaluate the visual odometry based on a focused plenoptic camera several experiments were performed. In these experiments we focused on measuring the accuracy of recorded objects in 3D space. Besides, we want to demonstrate the capabilities of the focused plenoptic camera based visual odometry in a sample scene. Since our interest lies in the depth map of the scene, we did not measure a ground truth for the performed trajectories and thus, the tracking itself was not evaluated explicitly.

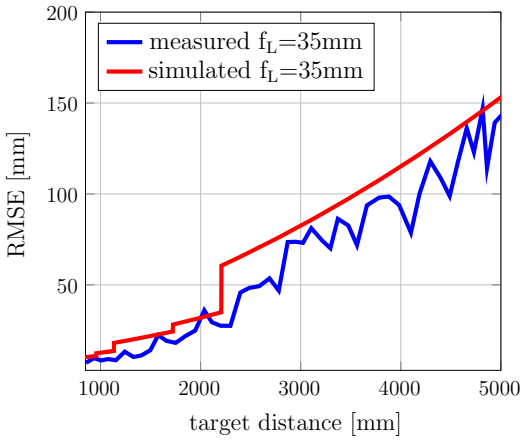


Fig. 15. Depth accuracy of a Raytrix R5 camera with main lens focal lengths $f_L = 35$ mm.

Fig. 16. Sample scene which was recorded to evaluate the focused plenoptic camera based visual odometry.

For the experiments presented in this section, we reduced the image resolution of both the totally focused intensity image as well as the corresponding virtual depth map, to 512 × 512 pixel.

7.3.1.1. 3D reconstruction. For a qualitative evaluation of our method, we recoded an image sequence composed of 2600 frames by the focused plenoptic camera. Fig. 16 shows the recorded scene and Fig. 17 some sample images out of the recorded sequence. The shown sequence was recorded with a frame rate of approximately 50 fps while the camera was moved freehand. Thus, between two consecutive frames on average a translation of roughly 2 mm is performed. After recording, our focused plenoptic camera based visual odometry algorithm is applied offline to the sequence.

7.3.1.2. Depth accuracy. Two experiments were performed to evaluate the depth accuracy of the focused plenoptic camera based visual odometry. We evaluate the depth over time as well as over object distance. For both experiments the same setup as for the depth calibration (Section 7.2.1) was used.

To see how the visual odometry improves the depth information, an image sequence is recorded while the camera is translated in vertical direction. For each object distance a vertical movement of 20 cm was performed while recording the image sequence.

In the first experiment the depth accuracy over a sequence of images, while the camera is moving in vertical direction, is evaluated for object distances from approx. 2.6 m to 5.3 m with a spacing of 30 cm. Exemplary we present the results for an object distance of 3.183 m. The calculated metric depth map at each frame is read out and analyzed. Since the camera is moved more or less uniformly over time, this evaluation is equivalent to measuring the accuracy as function of baseline distance.

The second experiment is performed to evaluate the depth accuracy of the plenoptic camera based visual odometry with respect to the object distance a_L. Therefor the standard deviation of the depth, which resulted from the plenoptic camera based visual odometry, was evaluated for the 10 object distances in the range from 2.6 m to 5.3 m, after a vertical translation of 20 cm.

7.3.2. Results

7.3.2.1. 3D reconstruction. Fig. 18 shows the 3D point cloud which resulted from the focused plenoptic camera based visual odometry. Of course, this point cloud gives only a qualitative impression. Nevertheless, one can see, that for instance the rectangular shape of the table itself and other items on the table, like the keyboard or the book on the right are kept. Even though in this figure the point clouds which resulted from over 40 reference frames are overlaid no misalignment between them can be seen.

There is quite a big gap at the position where the computer monitor should be in the point cloud. The problem is that the monitor has almost no structure in the recorded sequence (see Fig. 17) and thus no reliable depth can be estimated.

Due to the very narrow FOV of about 18° for a monocular visual odometry no reliable tracking would be possible, as shown by Zeller et al. (2015b). This experiment shows that our plenoptic camera based approach is working in setups where monocular approaches fail and thereby extends the working range of visual odometry in general.

7.3.2.2. Depth accuracy. Fig. 19 shows the course of the depth's standard deviation over all frames in the recorded sequence and thus as a discrete function over time. Since the sequence was recorded for a more or less homogeneous movement in vertical direction, the standard deviation can also be considered as a function of the baseline distance to the first frame. As already mentioned, the baseline between the first and the last frame is 20 cm in length.

One can see from Fig. 19 that the curve has approximately $1/x$ behavior. This conforms to the theoretical depth accuracy of a pair of stereo images recorded for the simplified case. Here one can derive the depth accuracy based on the theory of propagation of uncertainty, as given in Eq. (73).

Fig. 17. Five intensity image samples out of the sequence of 2600 frames recorded by a Raytrix R5 camera.

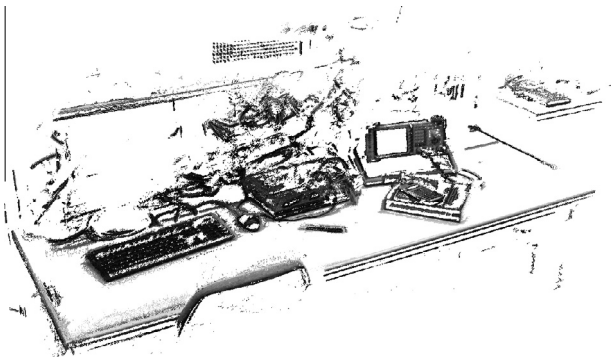


Fig. 18. 3D point clouds of a sample scene recorded by a Raytrix R5 camera after applying the focused plenoptic camera based visual odometry to a sequence of 2600 frames.

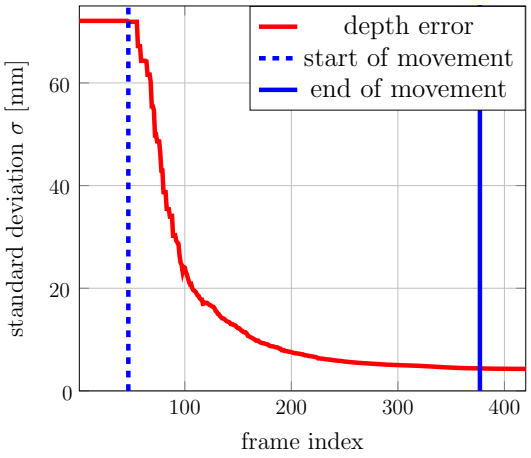


Fig. 19. Standard deviation of the measured depth for a chessboard target in 3.183 m distance to the camera. From the first to the last frame the camera was translated by 20 cm in vertical direction.

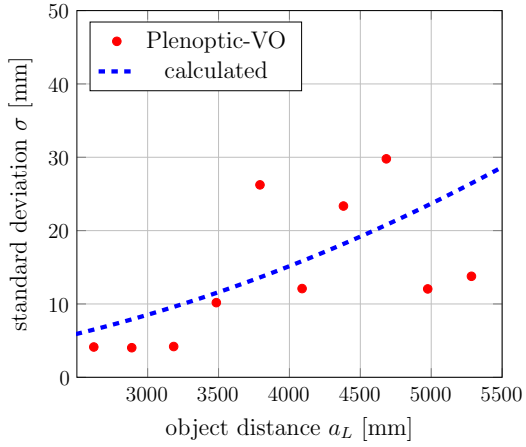


Fig. 20. Standard deviation measured over object distance. Red dots: Standard deviation of the focused plenoptic camera based visual odometry after a translation of 20 cm in vertical direction. Blue dashed line: Standard deviation for a stereo camera pair with baseline distance of 20 cm, intrinsic parameters similar to the Raytrix camera, at a disparity standard deviation of 0.3 pixel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\sigma_{a_L} = \frac{f \cdot d_B}{p_x^2} \cdot \sigma_{p_x} = \frac{a_L^2}{f \cdot d_B} \cdot \sigma_{p_x} \tag{73}$$

In Eq. (73) f represents the focal length of the underlying pinhole camera model, d_B the baseline distance, a_L the object distance and p_x the measured disparity. The standard deviation of the disparity σ_{p_x} can be considered as constant.

After the camera started moving there is still a range of about seven frames where the standard deviation does not decay. In this range the baseline to the first frame is too short to improve the depth of the focused plenoptic camera and thus, no improvement in the depth accuracy is achieved. Thereafter, the larger baseline built by subsequent frames leads to a quite steep descent of the depth's standard deviation.

Fig. 20 shows the results for the chessboard plane recorded for different object distances a_L. Here the red dots show the depth's standard deviation after the focused plenoptic camera based visual odometry with a translation of 20 cm.

The blue dashed line represents the theoretical depth standard deviation for a stereo camera pair with a baseline distance of 20 cm, intrinsic parameters similar to those of the Raytrix camera, and a disparity standard deviation σ_{p_x} of 0.3 pixel. This curve can be calculated from Eq. (73). Thus, the measured values conform to the theoretical limits.

8. Summary and conclusion

In this paper we addressed the problem of depth estimation for a focused plenoptic camera and presented a probabilistic depth estimation approach. We showed how the camera can be calibrated to be used in photogrammetric or computer vision applications and additionally we were able to run a visual odometry algorithm based on a focused plenoptic camera. This approach has certain advantages compared to a monocular visual odometry.

For the proposed depth estimation algorithm we introduce a graph of baselines which defines the multiple micro lens pairs in the MLA. Based on this graph multiple binocular stereo-observations are obtained, starting from a short up to a long baseline. These observations are incorporated in a probabilistic depth map. We expressed mathematically how the camera noise affects the disparity estimation. Thus, the estimated inverse virtual depths can be defined as Gaussian distributed random variables. Based on the probabilistic depth map it is possible to remove outliers without any low-pass filtering by setting a threshold for the inverse virtual depth variance. Thus, discontinuities in the depth map are preserved.

For camera calibration we apply a traditional camera model to the synthesized image of a focused plenoptic camera and estimate it based on a traditional calibration method. We developed two model based depth calibration methods, which proved to define the camera model very well, and compared them to a well known curve fitting approach. Due to the precise models, only a small number of measurements is needed for calibration. Besides, it was shown that the estimated functions are valid in excess of the calibration range.

In this paper we incorporated both, the estimated depth map and the camera model, which resulted from calibration, into a focused plenoptic camera based visual odometry. We achieve considerable improvements both with respect to the depth from a single image of the focused plenoptic camera and the depth received from a monocular visual odometry. We were able to run the algorithm when using a focused plenoptic camera with a narrow FOV and a respectively large focal length. Another main improvement is that our focused plenoptic camera based visual odometry also measures scale and thus metric tracking and mapping is possible.

Compared to a monocular camera, the hardware effort stays pretty much the same, since the plenoptic camera differs only by the MLA in front of the sensor. The computational effort however

increases compared to monocular visual odometry since light-field based depth estimation has to be performed.

Acknowledgement

This research is funded by the Federal Ministry of Education and Research of Germany in its program "IKT 2020 Research for Innovation".

References

Adelson, E.H., Wang, J.Y.A., 1992. Single lens stereo with a plenoptic camera. IEEE Trans. Pattern Anal. Machine Intell. 14, 99–106 http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15869-f11/www/readings/adelson92_plenoptic.pdf.

Akbarzadeh, A., Frahm, J.M., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Yang, R., Welch, G., Towles, H., Nister, D., Pollefeys, M., 2006. Towards urban 3d reconstruction from video. In: Proc. Third International Symposium on 3D Data Processing, Visualization, and Transmission, pp. 1–8. http://dx.doi.org/10.1109/3DPVT.2006.141 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4155703&tag=1.

Bishop, T.E., Favaro, P., 2011. Full-resolution depth map estimation from an aliased plenoptic light field. In: Kimmel, R., Klette, R., Sugimoto, A. (Eds.), Computer Vision – ACCV 2010, Lecture Notes in Computer Science, vol. 6493. Springer, Berlin Heidelberg, pp. 186–200. http://dx.doi.org/10.1007/978-3-642-19309-5_15 http://link.springer.com/chapter/10.1007%2F978-3-642-19309-5_15.

Brown, D.C., 1966. Decentering distortion of lenses. Photogr. Eng. 32, 444–462 https://eserv.asprs.org/PERS/1966journal/may/1966_may_444-462.pdf.

Concha, A., Civera, J., 2014. Using superpixels in monocular slam. In: Proc. IEEE International Conference on Robotics and Automation (ICRA), pp. 365–372. http://dx.doi.org/10.1109/ICRA.2014.6906883 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6906883&tag=1.

Dansereau, D., Mahon, I., Pizarro, O., Williams, S., 2011. Plenoptic flow: closed-form visual odometry for light field cameras. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4455–4462. http://dx.doi.org/10.1109/IROS.2011.6095080 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6095080.

Dansereau, D., Pizarro, O., Williams, S., 2013. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1027–1034. http://dx.doi.org/10.1109/CVPR.2013.137 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6618981&tag=1.

Eade, E., Drummond, T., 2009. Edge landmarks in monocular slam. Image Vision Comput. 27, 588–596. http://dx.doi.org/10.1016/j.imavis.2008.04.012 http://www.sciencedirect.com/science/article/pii/S0262885608000978#.

Engel, J., Schöps, T., Cremers, D., 2014. Lsd-slam: large-scale direct monocular slam. In: Proc. European Conference on Computer Vision (ECCV), pp. 834–849 http://link.springer.com/book/10.1007/978-3-319-10605-2/page/3.

Engel, J., Sturm, J., Cremers, D., 2013. Semi-dense visual odometry for a monocular camera. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 1449–1456. http://dx.doi.org/10.1109/ICCV.2013.183 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6751290.

Forster, C., Pizzoli, M., Scaramuzza, D., 2014. Svo: fast semi-direct monocular visual odometry. In: Proc. IEEE International Conference on Robotics and Automation (ICRA), pp. 15–22. http://dx.doi.org/10.1109/ICRA.2014.6906584 http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6906584&queryText%3DSVO%3A+Fast+Semi-Direct+Monocular+Visual+Odometry.

Georgiev, T., Lumsdaine, A., 2009. Depth of field in plenoptic cameras. In: Eurographics http://www.tgeorgiev.net/DepthOfField.pdf, short paper.

Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F., 1996. The lumigraph. In: Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH. ACM, New York, NY, USA, pp. 43–54. http://dx.doi.org/10.1145/237170.237200 http://dl.acm.org/citation.cfm?doid=237170.237200.

Heber, S., Pock, T., 2014. Shape from light field meets robust pca. In: Proc. European Conference on Computer Vision (ECCV), pp. 751–767 http://link.springer.com/chapter/10.1007%2F978-3-319-10599-4_48.

Ives, F.E., 1903. Parallax Stereogram and Process of Making Same http://www.google.com/patents?id=ouBYAAAAEBAJ&printsec=abstract&zoom=4#v=onepage&q&f=false.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A., 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proc. 24th Annual ACM Symposium on User Interface Software and Technology. ACM, New York, NY, USA, pp. 559–568. http://dx.doi.org/10.1145/2047196.2047270 http://dl.acm.org/citation.cfm?doid=2047196.2047270.

Jeon, H.G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., Kweon, I.S., 2015. Accurate depth map estimation from a lenslet light field camera. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1547–1555.

Johannsen, O., Heinze, C., Goldluecke, B., Perwaß, C., 2013. On the calibration of focused plenoptic cameras. In: GCPR Workshop on Imaging New Modalities http://hci.iwr.uni-heidelberg.de//Staff/bgoldlue/.

Kerl, C., Sturm, J., Cremers, D., 2013. Dense visual slam for rgb-d cameras. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2100–2106. http://dx.doi.org/10.1109/IROS.2013.6696650 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6696650.

Kim, M.J., Oh, T.H., Kweon, I.S., 2014. Cost-aware depth map estimation for lytro camera. In: Proc. IEEE International Conference on Image Processing (ICIP), pp. 36–40. http://dx.doi.org/10.1109/ICIP.2014.7025006.

Klein, G., Murray, D., 2007. Parallel tracking and mapping for small ar workspaces. In: Proc. IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR), pp. 225–234. http://dx.doi.org/10.1109/ISMAR.2007.4538852 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4538852.

Li, M., Mourikis, A.I., 2013. High-precision, consistent ekf-based visual-inertial odometry. Int. J. Robot. Res. 32, 690–711. http://dx.doi.org/10.1177/0278364913481251 http://ijr.sagepub.com/content/32/6/690.abstract.

Lippmann, G., 1908. Epreuves reversibles donnant la sensation du relief. J. Phys. Théor. Appl. 7, 821–825 http://www.tgeorgiev.net/Lippmann/.

Lumsdaine, A., Georgiev, T., 2008. Full resolution lightfield rendering. Technical Report, Adobe Systems, Inc. <http://www.tgeorgiev.net/FullResolution.pdf>.

Lumsdaine, A., Georgiev, T., 2009. The focused plenoptic camera. In: Proc. IEEE International Conference on Computational Photography (ICCP), San Francisco, CA, pp. 1–8 http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=5559008.

Newcombe, R.A., Lovegrove, S.J., Davison, A.J., 2011. Dtam: dense tracking and mapping in real-time. In: Proc. IEEE International Conference on Computer Vision (ICCV) http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.6884.

Ng, R., 2006. Digital light field photography Ph.D. thesis. Stanford University, Stanford, USA.

Perwaß, C., Wietzke, L., 2012. Single lens 3d-camera with extended depth-of-field. In: Proc. SPIE 8291, Human Vision and Electronic Imaging XVII. Burlingame, California, USA http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1283425.

Schöps, T., Engel, J., Cremers, D., 2014. Semi-dense visual odometry for ar on a smartphone. In: Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 145–150. http://dx.doi.org/10.1109/ISMAR.2014.6948420 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6948420.

Tao, M.W., Hadap, S., Malik, J., Ramamoorthi, R., 2013. Depth from combining defocus and correspondence using light-field cameras. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 673–680. http://dx.doi.org/10.1109/ICCV.2013.89 http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6751193&queryText%DDepth+from+combining+defocus+and+correspondence+using+light-field+cameras.

Tosic, I., Berkner, K., 2014. Light field scale-depth space transform for dense depth estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 441–448. http://dx.doi.org/10.1109/CVPRW.2014.71 http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6910019&queryText%3DLight+field+scale-depth+space+transform+for+dense+depth+estimation.

Venkataraman, K., Lelescu, D., Duparre, J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R., Nayar, S., 2013. Picam: an ultra-thin high performance monolithic camera array. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 32, 1–13. http://dx.doi.org/10.1145/2508363.2508390 http://dl.acm.org/citation.cfm?doid=2508363.2508390.

Wanner, S., Goldluecke, B., 2012. Globally consistent depth labeling of 4d lightfields. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) http://hci.iwr.uni-heidelberg.de/HCI/Research/LightField/#publications.

Wanner, S., Goldluecke, B., 2014. Variation light field analysis for disparity estimation and super-resolution. IEEE Trans. Pattern Anal. Machine Intell. 36, 606–619.

Yu, Z., Guo, X., Lin, H., Lumsdaine, A., Yu, J., 2013. Line assisted light field triangulation and stereo matching. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2792–2799 http://www.cv-foundation.org/openaccess/content_iccv_2013/html/Yu_Line_Assisted_Light_2013_ICCV_paper.html.

Zeller, N., Quint, F., Stilla, U., 2014. Calibration and accuracy analysis of a focused plenoptic camera. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3, 205–212. http://dx.doi.org/10.5194/isprsannals-II-3-205-2014 http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/II-3/205/2014/.

Zeller, N., Quint, F., Stilla, U., 2015a. Establishing a probabilistic depth map from focused plenoptic cameras. In: Proc. International Conference on 3D Vision (3DV), pp. 91–99. http://dx.doi.org/10.1109/3DV.2015.18.

Zeller, N., Quint, F., Stilla, U., 2015b. Narrow field-of-view visual odometry based on a focused plenoptic camera. ISPRS Ann. Photogr. Remote Sens. Spatial Inform. Sci. II-3/W4, 285–292. http://dx.doi.org/10.5194/isprsannals-II-3-W4-285-2015 http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/II-3-W4/285/2015/.