Depth Range Accuracy for Plenoptic Cameras

Nuno Barroso Monteiro

Institute for Systems and Robotics, University of Lisbon, Portugal
Institute for Systems and Robotics, University of Coimbra, Portugal

Simão Marto

Institute for Systems and Robotics, University of Lisbon, Portugal

João Pedro Barreto

Institute for Systems and Robotics, University of Coimbra, Portugal

José Gaspar

Institute for Systems and Robotics, University of Lisbon, Portugal

Abstract

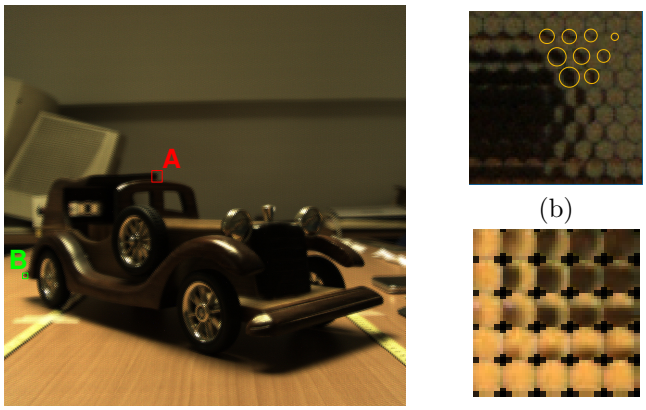
Plenoptic cameras capture the directional information of the light distribution from a scene. This is accomplished by positioning a microlens array between the main lens and the sensor. This configuration obtains multiple projections for a point in the object space, which allows to retrieve the point's depth on a single exposure. In recent years, several studies recover depth and shape from the lightfield data using several cues. Nonetheless, references regarding the depth capabilities of a standard plenoptic camera with different zoom and focus settings are scarce. In this work, we formalize a forward projection model and consider projection geometry cues to improve a metric reconstruction methodology for a calibrated standard plenoptic camera. The metric reconstruction methodology is used to evaluate the depth estimation accuracy under certain zoom and focus settings. The reconstruction is applied to new datasets captured for this purpose with objects placed at depths between 0.05 and 2.00 meters. The results indicate that these cameras are able to reconstruct accurately points within the depth range analyzed by appropriately choosing the zoom and focus depth settings. The zoom is a determinant factor on the reconstruction accuracy and the focus depth allows to determine the reconstruction depth range.

Keywords: Standard Plenoptic Camera, Projection, Reconstruction, Depth Range

Email addresses: nmonteiro@isr.tecnico.ulisboa.pt (Nuno Barroso Monteiro), smarto@isr.tecnico.ulisboa.pt (Simão Marto), jpbar@isr.uc.pt (João Pedro Barreto), jag@isr.tecnico.ulisboa.pt (José Gaspar)

1. Introduction

Images obtained using conventional cameras, such as the pinhole camera, capture the total amount of light that reaches each position of the sensor. In these cameras, a point in the object space is projected onto a single pixel. Plenoptic cameras, on the other hand, are capable of discriminating the contribution of each light ray emanating from a particular point by projecting the point to several positions of the sensor (orange circles in Figure 1.b).



(a) (b) (c)

Figure 1: **(a)** Image captured on the sensor of a standard plenoptic camera. This image has a resolution of 3280 x 3280 pixels. **(b)** Magnification of red box A in (a). This image depicts the microlenses images formed in the sensor. The orange circles show the multiple projections of the corner point of the upper cover of the car. The corner is projected onto several microlenses because it is beyond the plane of focus situated on the middle of the car. **(c)** Microlenses images obtained after decoding the raw image to the 4D lightfield. The microlenses correspond to the green box B highlighted in (a).

There are several types of plenoptic cameras like the artificial compound eyes [1, 2], the wavefront coding systems [3], or the lenticular array based [4, 5]. The lenticular array based plenoptic cameras consist of a main lens and microlens array. In this work, we will analyze the depth estimation accuracy for a lenticular array based plenoptic camera, more specifically, the standard plenoptic camera introduced by Ng [6]. The standard plenoptic camera [6] has a higher directional resolution and produces images with lower spatial resolution [7] when compared to the focused plenoptic camera introduced by Lumsdaine and Georgiev [8, 9].

Depth estimation is one of the applications found in the literature since these cameras allow to retrieve depth from a single exposure. Most of the depth reconstruction studies consider a camera geometry that resembles a camera array, usually a single camera mounted on a gantry system [10, 11, 12, 13], or using simulated environments [14, 12, 13, 15]. Nonetheless, the geometry of

plenoptic cameras is more complex. The number of studies recovering depth using plenoptic cameras is limited [16, 17, 12, 15] and even fewer using standard plenoptic cameras [15]. Additionally, references regarding the depth capabilities, i.e. the accuracy of the reconstructed depth of these sensors is also scarce. There are studies on the depth capabilities for a focused plenoptic camera [16, 17] but to the authors knowledge there are no similar studies for standard plenoptic cameras. The depth estimation and capabilities of these sensors depend on the world plane in focus by the main lens. Thus, to study these sensors a combination of camera parameters must be analyzed to assess the reconstruction estimation accuracy.

In this work, we define a forward projection model that establishes a relationship between a point in the object space and the lightfield on the sensor plane based on the camera model of Dansereau et al. [18]. Analyzing the projection model, we improve a reconstruction methodology by adding projection geometry cues. This reconstruction methodology is then used to evaluate the depth capabilities of a calibrated standard plenoptic camera for different zoom and focus settings. The results presented suggest that these cameras are able to reconstruct accurately points within the depth range analyzed by appropriately choosing the zoom and focus depth settings. Namely, zoom is a determinant factor on the reconstruction accuracy and the focus depth allows to determine the reconstruction depth range.

Contributions. The contributions of this work are three-fold: (i) definition of the geometry for standard plenoptic cameras, (ii) analysis of the depth capabilities of a standard plenoptic camera for different zoom and focus settings, (iii) and a database to calibrate and assess the reconstruction accuracy of a standard plenoptic camera for different zoom and focus settings.

In terms of structure, we present in Section 2 a brief review of camera models and depth estimation methods considered for plenoptic cameras. In Section 3, we introduce some concepts found in the literature to contextualize the reader: the lightfield parameterization, and the back-projection model introduced by Dansereau et al. [18]. The forward projection model is formalized and analyzed in Section 4 while the reconstruction methodologies are presented and compared in Section 5. The results of the reconstruction at different depths for a calibrated plenoptic camera under certain zoom and focus settings are presented in Section 6. The major conclusions are presented in Section 7.

Notation: The notation followed throughout this work is the following: non-italic letters correspond to functions, italic letters correspond to scalars, lower case bold letters correspond to vectors, and upper case bold letters correspond to matrices.

2. Related Work

Depth estimation was the first application to be studied with a plenoptic camera prototype [4]. This topic is still an active line of research. Nonetheless, studies regarding the depth capabilities of these sensors are scarce. In this

section, we will highlight some studies regarding the properties of plenoptic cameras and some methods for depth estimation from the lightfield.

2.1. Plenoptic Camera Studies

The camera models defined for the standard plenoptic camera are very scarce. Dansereau et al. [18] obtained a camera model by tracing the rays from the sensor to the object space. In this model, Dansereau et al. [18] assumed that all light rays reaching a given pixel of the sensor cross the center of the microlens (pinhole model), and that the corresponding microlens is the one whose center gets projected closest to the pixel. The main lens was modeled as a thin lens. Nonetheless, Dansereau et al. [18] did not establish a relation of the lightfield with a point in the object space. The relationship between an arbitrary lightfield and a point in the object space appears in a previous work of Dansereau et al. [10]. There is another camera model proposed by Johannsen et al. [19] that considered Plücker coordinates to obtain the projection of a point on the lightfield in the object space. Nonetheless, these projections are not related with the lightfield in the sensor plane.

The depth capabilities studies for plenoptic cameras is also scarce. Recently, Johannssen et al. [16] and Zeller et al. [17] proposed calibration methods to improve the depth accuracy of focused plenoptic cameras. Nonetheless, there are no similar studies for standard plenoptic cameras. The most similar studies for standard plenoptic cameras correspond to the works of Hahne et al. [20, 21]. These studies estimated depth, depth of field and baselines using different optical parameters for the microlenses and for the main lens of a simulated and a customized standard plenoptic camera. Nonetheless, these works require the specific knowledge of the parameters of the optical setup and are more focused on assisting the specification to design a standard plenoptic camera.

2.2. Depth Estimation

Recent approaches on depth estimation from the lightfield [10, 11, 12, 13, 14] consider the epipolar plane images (EPIs) geometry [22]. These works consider a lightfield acquired using a camera geometry that resembles a camera array. According to Bolles et al. [22], the slope of the lines found in the EPIs relate to the depth of a point in the object space. Hence, the point correspondence problem is now a problem of finding lines in the EPI. Dansereau et al. [10] estimated the slopes of the lines in the EPIs using image gradients while Wanner et al. [12] uses a structure tensor analysis. In Wanner et al. [12] a fast denoising method is used to obtain a dense disparity map from a limited region of the full 4D lightfield. Monteiro et al. [14] considered a different optimization scheme (SALSA [23]) to integrate the sparse estimates. The performance is improved by considering periodic boundary conditions. These methods do not handle occlusion explicitly and have a small disparity range. Diebold et al. [11] increased the allowed disparity range by shearing the lightfield, and handled occlusion by integrating the estimates using a specific metric. Recently, Lüke et al. [13] proposed a method for depth estimation based on the local gradient

of the 4D lightfield for a given ray. Other approaches consider virtual cameras, the surface cameras (SCams) [24], that collect rays of the lightfield intersecting at a point in the object space. These collections of rays can be used to identify correspondences, detect occlusions or surface characteristics [24, 25]. There are other methods for depth estimation from lightfield, for more detail please refer to Johannsen et al. [26]. Nonetheless, the setup of a standard plenoptic camera, namely the geometry introduced by the main lens and the microlens array, introduces a more complex geometry that still requires studies assessing metric reconstruction.

Regarding standard plenoptic cameras, Ng et al. [6] described a different approach that allows to recover relative non-metric depth from defocus. More recently, Tao et al. [15] proposed to unify different cues obtained from the EPI (defocus and correspondence) to estimate depth. As in the work of Ng et al. [6], this estimate corresponds to a relative depth and not absolute depth. The depth is given relatively to the world focal plane. In a more recent work, Tao et al. [27] included shading to recover shape. By considering the shading cues, the method improved shape estimation for specular surfaces.

3. Lightfield Parameterization and Back-Projection

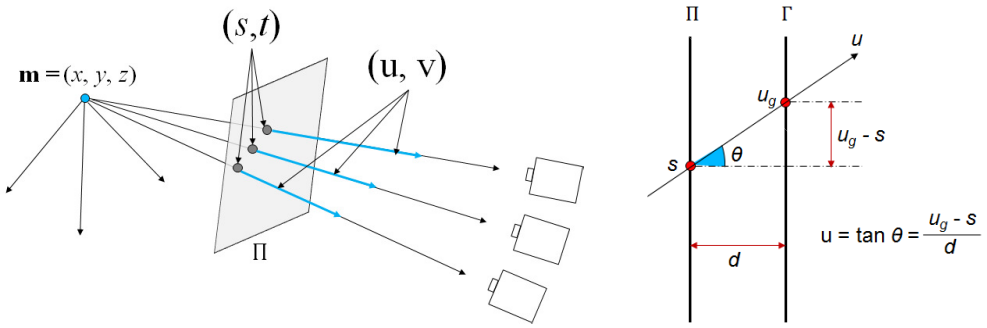
In this section, we will highlight some key aspects found in the literature that are relevant to the contributions presented in Sections 4 and 5. Namely, we will briefly detail the lightfield parameterization, and the back-projection model presented by Dansereau et al. [18].

3.1. Two-Plane Parameterization

The lightfield [28, 29] is a simplification of the plenoptic function [30] that considers static monochromatic light rays with constant radiance along the ray. The lightfield can be described using a two-plane parameterization which represent a ray by its intersection with two planes.

In this work, we use a two points representation [31] to define the lightfield (i, j, k, l) on the sensor plane while a point and a direction representation is used to define the lightfield (s, t, u, v) in the object space. In the object space, (s, t) corresponds to a point in an arbitrary plane and (u, v) to a ray direction (Figure 2). While in the sensor plane, (i, j) corresponds to the point in the image sensor defined in the local reference frame of the corresponding microlens and (k, l) corresponds to the point in the microlens plane. The parameterization with a point and a direction is equivalent to a local two-plane parameterization using two points considering a unitary distance between the two planes.

To retrieve the two-plane parameterization from the image sensor (raw image in Figure 1.a) a decoding process is necessary, like the one introduced by Dansereau et al. [18]. The decoding process consists on defining the pixels on the image sensor that belong to a given microlens. The pixels that belong to a given microlens are considered to be the ones nearest to the projection of the microlens center on the image sensor. For more detail on the decoding process please refer to [18].



(a) (b)

Figure 2: **(a)** Two plane parameterization for rays starting from a point $\mathbf{m}$ using a point and a direction. Each of these rays correspond to different views of a point, i.e., they are samples of the lightfield. **(b)** The intersection of a ray with the two planes (s,t) and (u_g, v_g) define a geometry that allows to determine the direction of the ray. The plane Π defines the position of the rays while the plane Γ defines the direction of the rays.

3.2. Back-Projection

Consider that we have an arbitrary point $\mathbf{m} = [x, y, z]^T$. Note that $\mathbf{m}$ is defined in the camera coordinate system. Let us define a plane Π that comprises the origin of the different rays $\mathbf{\Psi} = [s, t, u, v, 1]^T$ in the object space. The center of this plane corresponds to the origin of the camera coordinate system. Using the different positions (s,t) of the rays in this plane and their directions (u,v), the relation between a point and the lightfield in the object space is defined as in the work of Grossberg and Nayar [32]

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s \\ t \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad , \; \lambda \in \mathbb{R} \quad , \tag{1}$$

with $u = \frac{x-s}{z}$ and $v = \frac{y-t}{z}$. This equation allows to propagate the position in the rays originated at $\mathbf{m}$ to an arbitrary plane at distance λ from the origin of the camera coordinate system. Note that equation (1) generalizes a normalized pin-hole camera: by setting $s = 0$, $t = 0$ and $(u,v) \in \mathbb{R}^2$ one obtains a pencil of lines. Therefore, by allowing $(s,t) \in \mathbb{R}^2$, one can represent an infinite number of normalized pin-hole cameras.

Now, let us consider the relation between the lightfield in the sensor plane with the lightfield in the object space. The model proposed by Dansereau et al. [18] allows to map the lightfield on the sensor plane $\mathbf{\Phi} = [i, j, k, l, 1]^T$ to the lightfield in the object space $\mathbf{\Psi}_g = [s, t, u_g, v_g, 1]^T$ using the matrix $\mathbf{H}_g$. $\mathbf{\Psi}_g$ defines a lightfield using a two-plane parameterization with two points, where (u_g, v_g) correspond to the intersection of the ray with a plane Γ parallel to and at a distance d from the plane Π. (u_g, v_g) are global coordinates defined

relatively to the origin of the camera coordinate system. Thus, the lightfield in the object space $\mathbf{\Psi} = [s, t, u, v, 1]^T$ defined using a two-plane parameterization with a point and a direction can be obtained considering the geometry presented in Figure 2.b. Ultimately, this leads to the following relationship between the lightfields in the sensor plane and in the object space

$$\underbrace{\begin{bmatrix} s \\ t \\ u \\ v \\ 1 \end{bmatrix}}_{\mathbf{\Psi}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1/d & 0 & 1/d & 0 & 0 \\ 0 & -1/d & 0 & 1/d & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{H}_g}_{\mathbf{H}} \underbrace{\begin{bmatrix} i \\ j \\ k \\ l \\ 1 \end{bmatrix}}_{\mathbf{\Phi}} . \tag{2}$$

The distance between the planes Π and Γ in the object space is set arbitrarily. Hence, throughout the remainder of the work we will consider the distance equal to one (d = 1). The lightfield on the sensor plane corresponds to a virtual lightfield obtained after the decoding process in [18]. This lightfield follows the same notation of Dansereau et al. [18], that considered (k, l) as the indices of the microlenses and (i, j) as the relative indices of the pixels within each microlens. The matrix $\mathbf{H}$ allows to map rays defined in pixels and microlenses indices to rays defined by a position and a direction in metric units. $\mathbf{H}$ is a matrix containing intrinsic parameters

$$\mathbf{H} = \begin{bmatrix} h_{si} & 0 & h_{sk} & 0 & h_s \\ 0 & h_{tj} & 0 & h_{tl} & h_t \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} . \tag{3}$$

The lightfield in the object space, $\mathbf{\Psi}$, is defined on the plane Π (Figure 2.a). The lightfield obtained in the object space is characterized by rays whose directions have not been modified by the camera optics. The model presented by Dansereau et al. [18] is not explicitly a projection model in the sense that it does not relate the rays $\mathbf{\Psi}$ with a specific point in the object space. Combining this model with a point using equation (1), one obtains a back-projection model from image to object space.

3.2.1. Radial Distortion

The camera model proposed by Dansereau et al. [18] considered a radial distortion model to compensate for the lens distortion caused by the microlens and main lens optics that cannot be described using equation (2). This model compensates for the lens distortion by modifying the coordinates (u_g, v_g) of the lightfield in the object space. For the parameterization using a point and a direction the radial distortion compensates the distortion modifying the directions (u, v) of the rays in the object space according to the distance to a distortion

center that defines the central ray $\mathbf{r}_{uv}$. Let $\mathbf{\Psi}^u_{uv}$ denote the undistorted directions (u, v), the distorted directions $\mathbf{\Psi}^d_{uv}$ are given by

$$\mathbf{\Psi}^d_{uv} = \left(1 + \sum_{n=1}^{M} k_n \, r^{2\,n}\right) (\mathbf{\Psi}^u_{uv} - \mathbf{r}_{uv}) + \mathbf{r}_{uv} \tag{4}$$

where $r = \|\mathbf{\Psi}^u_{uv} - \mathbf{r}_{uv}\|$ is the distance from a given ray (u, v) to the central ray, $k_1, \ldots, k_M$ are the radial distortion coefficients, and M is the number of coefficients considered to model the radial distortion. In this work, we will consider $M = 4$ radial distortion coefficients. The distortion center is considered to be different from the optical axis of the camera. For details on how to rectify the lightfield coordinates refer to [18]. For the following sections consider that whenever $\mathbf{\Psi}$ or $\mathbf{\Phi}$ are represented without superscripts, we are considering undistorted coordinates.

Let us introduce additional concepts found in the literature which are going to be necessary in the following sections. The lightfield on the sensor plane allows to define two types of image representations, namely microlens and viewpoint images [6]. A microlens image (Figure 1.c) results from the rays that cross the center of a specific microlens, i.e., by fixing the microlens coordinates (k, l). A viewpoint or sub-aperture image (Figure 6.d) is obtained by selecting and combining the rays that reach the same pixel of each microlens, i.e, by selecting the pixel (i, j) of each microlens (k, l). In this case, the coordinates (i, j) are the indices of each viewpoint image and the coordinates (k, l) encodes the position of the pixel in the viewpoint image.

4. Projection Model

The projection model maps an arbitrary point in the object space, $\mathbf{m} = [x, y, z]^T$, to the lightfield on the sensor plane, $\mathbf{\Phi}$, knowing the intrinsic matrix $\mathbf{H}$. Rewriting equation (1) as $[x, y]^T = [s, t]^T + z\,[u, v]^T$ and replacing the lightfield in the object space $\mathbf{\Psi}$ with the lightfield on the sensor plane $\mathbf{\Phi}$ using the mapping defined in equation (2), one obtains:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{H}^{st}_{ij} \begin{bmatrix} i \\ j \end{bmatrix} + \mathbf{H}^{st}_{kl} \begin{bmatrix} k \\ l \end{bmatrix} + \mathbf{h}_{st} + z \, \left(\mathbf{H}^{uv}_{ij} \begin{bmatrix} i \\ j \end{bmatrix} + \mathbf{H}^{uv}_{kl} \begin{bmatrix} k \\ l \end{bmatrix} + \mathbf{h}_{uv} \right) \quad , \tag{5}$$

where the intrinsic matrix $\mathbf{H}$ is partitioned in four 2×2 diagonal sub-matrices

$$\begin{aligned} \mathbf{H}^{st}_{ij} &= \begin{bmatrix} h_{si} & 0 \\ 0 & h_{tj} \end{bmatrix}, \; \mathbf{H}^{st}_{kl} = \begin{bmatrix} h_{sk} & 0 \\ 0 & h_{tl} \end{bmatrix}, \\ \mathbf{H}^{uv}_{ij} &= \begin{bmatrix} h_{ui} & 0 \\ 0 & h_{vj} \end{bmatrix}, \; \mathbf{H}^{uv}_{kl} = \begin{bmatrix} h_{uk} & 0 \\ 0 & h_{vl} \end{bmatrix} \end{aligned} \quad , \tag{6}$$

and two 2×1 vectors $\mathbf{h}_{st} = [h_s, h_t]^T$ and $\mathbf{h}_{uv} = [h_u, h_v]^T$. Rewriting these equations relatively to (k, l), one can represent the projection model for a point $\mathbf{m}$ by

$$\begin{bmatrix} k \\ l \end{bmatrix} = \begin{bmatrix} \mathrm{f}\,(i; \mathbf{m}, \mathbf{H}) \\ \\ \mathrm{g}\,(j; \mathbf{m}, \mathbf{H}) \end{bmatrix} = \begin{bmatrix} -i\ \frac{h_{si} + z\,h_{ui}}{h_{sk} + z\,h_{uk}} + \frac{x - h_s - z\,h_u}{h_{sk} + z\,h_{uk}} \\ \\ -j\ \frac{h_{tj} + z\,h_{vj}}{h_{tl} + z\,h_{vl}} + \frac{y - h_t - z\,h_v}{h_{tl} + z\,h_{vl}} \end{bmatrix} \quad . \tag{7}$$

Note that $\mathrm{f}\,(i; \mathbf{m}, \mathbf{H})$ and $\mathrm{g}\,(j; \mathbf{m}, \mathbf{H})$ are mappings from $\mathbb{R} \to \mathbb{R}$, affine on the variables i and j. Since the point $\mathbf{m} \in \mathbb{R}^3$ and the intrinsic matrix $\mathbf{H} \in \mathbb{R}^5 \times \mathbb{R}^5$, the coordinates of the lightfield in the sensor plane (i, j, k, l), in general cannot be all integers. Equation (7) shows that a point in the object space defines lines on the spaces defined by each pair of coordinates (i, k) and (j, l). The space defined by these pair of coordinates is called the ray-space. These equations consider that the point $\mathbf{m}$ is defined in the camera coordinate system.

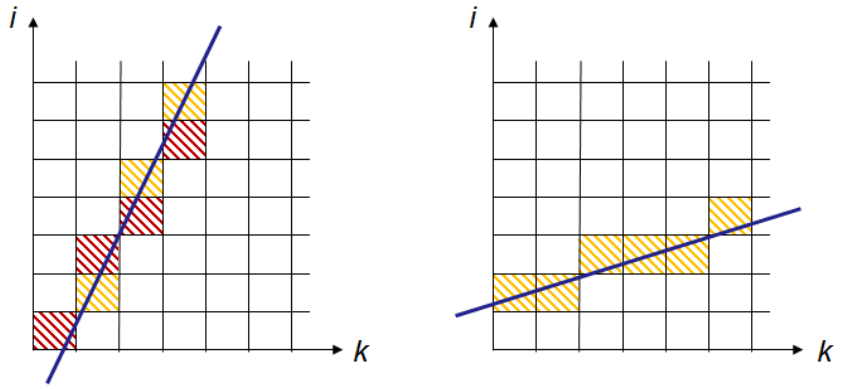
Unlike common projection problems, as in the pinhole camera model, in a standard plenoptic camera a point $\mathbf{m}$ in the object space can have multiple projections. In other words, the camera samples rays of the plenoptic function by having multiple projection centers. Thus, we want to maximize the number of projections obtained from the projection model.

4.1. Set of Imaged Points

A point in the object space projects into a line in the ray-space (i, k) and (j, l) (black line in Figure 3). The projection defined in equation (7) has 4 unknowns (i, j, k, and l) and 2 equations, which is not enough to define the rays $\mathbf{\Phi}$ on the sensor plane without any knowledge of the lightfield. Thus, we assume that the lightfield size is known. In a real camera one has a finite lightfield size that implies a finite number of rays $\mathbf{\Phi}$ obtained for the projection of a point $\mathbf{m}$.

Using the lightfield size and considering the discretization that occurs at the image sensor, we can assume integer values for the microlenses and determine the corresponding pixels. Nonetheless, according to the slope of the projection lines we can skip some projections since we are restricting the coordinates k, and l to be integers (red pixels in Figure 3.a). The same occurs if we assume integer values for the pixels and determine the corresponding microlenses. Since we want to maximize the number of projections, we evaluate the slope of the projection lines to determine which coordinates are more discriminative, the pixels or the microlenses.

Considering the linear mappings $k = \mathrm{f}\,(i; \mathbf{m}, \mathbf{H}) = m_k\ i + b_k$ and $l = \mathrm{g}\,(j; \mathbf{m}, \mathbf{H}) = m_l\ j + b_l$, the slope of the projection lines $m_{(\cdot)}$ corresponds to the disparity between viewpoint images, and its inverse corresponds to the disparity between microlens images. Slope $m_{(\cdot)}$ can also be identified in equation (7) as the factor multiplying i or j. Notice that the slope is constant for points at the same depth. $b_{(\cdot)}$ is the k- or l-intercept. To simplify our presentation, let us consider that the optical setup is point symmetric, i.e. the setup has square pixels and equally spaced microlenses in both vertical and horizontal directions.



(a) (b)

Figure 3: Rasterization method used to obtain the projections of a point $\mathbf{m}$ for the (i,k) coordinates for different slopes of the projection line $i = \mathrm{f}^{-1}(k;\mathbf{m},\mathbf{H})$. The red pixels correspond to the projections skipped by assuming integer values for the microlenses k.

This implies that $\mathrm{f}(i;\mathbf{m},\mathbf{H}) \equiv \mathrm{g}(j;\mathbf{m},\mathbf{H})$. Hence, if $\left|m_{(\cdot)}\right| \leq 1$, the pixels are more discriminative (Figure 3.a) and the microlens are given by the set $\mathcal{P}_{kl}$

$$\left\{ [i,j,k,l,1]^T : \; k = \mathrm{f}(i;\mathbf{m},\mathbf{H}),\; l = \mathrm{g}(j;\mathbf{m},\mathbf{H}),\; i \in \mathbb{N}_i,\; j \in \mathbb{N}_j \right\} \tag{8}$$

where $\mathbb{N}_i = \{0,\ldots,N_i-1\} \subset \mathbb{N}_0^+$, $\mathbb{N}_j = \{0,\ldots,N_j-1\} \subset \mathbb{N}_0^+$, and N_i and N_j correspond to the number of pixels of the sensor in each of the dimensions i and j. This is the case where a point $\mathbf{m}$ projects to more than one pixel within each microlens. This occurs, for example, if the point in the object space is near the focal plane or in focus by the main lens (Figure 1.b).

On the other hand, if $\left|m_{(\cdot)}\right| > 1$, the microlenses are more discriminative (Figure 3.b) and the pixels are given by the set $\mathcal{P}_{ij}$

$$\left\{ [i,j,k,l,1]^T : \; i = \mathrm{f}^{-1}(k;\mathbf{m},\mathbf{H}),\; j = \mathrm{g}^{-1}(l;\mathbf{m},\mathbf{H}),\; k \in \mathbb{N}_k,\; l \in \mathbb{N}_l \right\} \tag{9}$$

where $\mathbb{N}_k = \{0,\ldots,N_k-1\} \subset \mathbb{N}_0^+$, $\mathbb{N}_l = \{0,\ldots,N_l-1\} \subset \mathbb{N}_0^+$, and N_k and N_l correspond to the number of microlenses in each of the dimensions k and l. Since the camera might deviate from this point symmetric behavior, there might be some cases when we have to consider a correction using a mixture of the sets $\mathcal{P}_{ij}$ and $\mathcal{P}_{kl}$. For example, by considering $k = \mathrm{f}(i;\mathbf{m},\mathbf{H})$ and $j = \mathrm{g}^{-1}(l;\mathbf{m},\mathbf{H})$. The sets $\mathcal{P}_{ij}$ and $\mathcal{P}_{kl}$ describe a rasterization method for representing the lines defined in equation (7) for each of the coordinate pairs (i,k) and (j,l) in terms of discretized indices for pixels and microlenses. This process allows to implicitly overcome the limitations, detailed in Section 4.2, of the projection equation (7).

The complete projection model comprises the two sets $\mathcal{P}_{ij}$ and $\mathcal{P}_{kl}$, nonetheless, for most depth values the projection rays are obtained using the set $\mathcal{P}_{kl}$. The set $\mathcal{P}_{ij}$ is only used for points near the camera (see Supplementary Material).

4.2. Analysis of Singularities

The projection equation (7) has singularities. These singularities imply that some points in the object space have undefined projection rays (unobserved in the image). More precisely, $i = \mathrm{f}^{-1}(k; \mathbf{m}, \mathbf{H})$ or $k = \mathrm{f}(i; \mathbf{m}, \mathbf{H})$ are infinite for some depth values z, continuing with the point symmetric assumption.

The depth values for which the singularities occur are identified by $z_s^1 = -h_{si}/h_{ui}$ and $z_s^2 = -h_{sk}/h_{uk}$. Extending the definition of the entries h_{si}, h_{ui}, h_{sk}, and h_{uk} to consider the parameters proposed by Dansereau et al. [18] for the camera model, the singularities occur at $z_s^1 = \frac{d_M\, f_M}{d_M - f_M}$, and $z_s^2 = \frac{d_M\, f_M\, (F_s - N\, F_u) + F_s\, d_\mu\, f_M}{(d_M - f_M)\, (F_s - N\, F_u) + F_s\, d_\mu}$. Where N is the number of pixels in one dimension for the microlens image, d_M is the distance between the microlens plane and the main lens, d_μ is the focal length of the microlenses, and f_M is the focal length of the main lens. F_s and F_u are the spatial and directional sampling frequencies.

Looking more deeply into the singularities z_s^1 and z_s^2, we can see that z_s^1 corresponds to points that lie on the focal plane of the main lens. This can be derived from the thin lens equation for the main lens and remembering that the intrinsic matrix $\mathbf{H}$ propagates the origin of a ray to a plane that corresponds to the main lens plane. The depth of the singularity z_s^1 corresponds to the plane containing the projection centers of the microlens cameras. This singularity occurs when we apply the linear mapping $\mathrm{f}^{-1}(k; \mathbf{m}, \mathbf{H})$. Implicitly, the singularity implies that the slope m_k^{-1} is undefined. Thus, in Section 4.1, the set $\mathcal{P}_{kl}$ allows to overcome this limitation. This correction considers that a microlens can be defined by a range instead of an infinitesimal point which contradicts the initial assumption that the microlenses are pinholes.

On the other hand, the singularity z_s^2 corresponds to the depth of the plane containing the projection centers of the viewpoint cameras. The depth of the singularity is defined by the optical setup of the plenoptic camera and depends on several parameters including the sampling frequencies (see inline equation for z_s^2). This singularity occurs when we apply the linear mapping $\mathrm{f}(i; \mathbf{m}, \mathbf{H})$. Implicitly, the singularity implies that the slope m_k is undefined. However, when this situation occurs, we use the set $\mathcal{P}_{ij}$. This restriction avoids this limitation to occur in the projection model defined.

From these analysis, one can see that contrarily to a pinhole camera, a plenoptic camera can have projections even for points in the object space that are at the depths of the singularities z_s^1 and z_s^2.

Most of the studies in the literature recover depth from the lightfield by assuming the geometry of a camera array [10, 33, 13]. The common geometry for the camera array consists of identical parallel cameras with projection centers forming a regular planar grid. This geometry allows to relate the depth z

and the disparity Δu by $\Delta u = (f/z)\,\Delta s$, where Δs corresponds to a baseline measurement, and f to the focal length of the cameras.

The geometry of a camera array cannot be extended simply to a standard plenoptic camera. Let us consider the particular case of a point in focus by the main lens of the plenoptic camera. This point is represented by a vertical line in the ray-spaces (i, k) and (j, l), leading to projections onto one particular microlens. Therefore, in the viewpoint images this point in the object space will appear at exactly the same pixel position, i.e., the disparity for this point will be zero. Considering the geometry of the camera array, the point appears at infinity. However, for a plenoptic camera, this point has a well defined depth. The depth corresponds to the depth of the world focal plane of the main lens given by z_s^1 defined in Section 4.2. Thus, a more realistic scenario for the geometry of a plenoptic camera is a vergent camera array or a rectified parallel camera array.

4.3. Summary

In Section 4, we presented a forward projection model defined by the sets $\mathcal{P}_{ij}$ and $\mathcal{P}_{kl}$. This projection model was derived from the projection equation (7) considering the linear mappings $\mathrm{f}\,(i; \mathbf{m}, \mathbf{H})$ and $\mathrm{g}\,(j; \mathbf{m}, \mathbf{H})$ and the goal of maximizing the number of projections. In summary, the projection model can be defined using Algorithm 1. For simplicity, we presented the algorithm assuming that the optical system is point symmetric. C corresponds to the number of projection rays obtained.

Algorithm 1: Project scene point $\mathbf{m}$

Input : Scene point: $\mathbf{m} = \begin{bmatrix} x & y & z \end{bmatrix}^T$
Parameters: $\mathbf{H},\ N_i,\ N_j,\ N_k,\ N_l$
Output: Projection Rays: $\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_C\}$

Compute the slope m_i from equation (7)
if $|m_k| \leq 1$ **then**
  Rasterize $\mathbf{\Phi}_i = (i,\ j,\ k,\ l)$ according to set $\mathcal{P}_{kl}$;
else
  Rasterize $\mathbf{\Phi}_i = (i,\ j,\ k,\ l)$ according to set $\mathcal{P}_{ij}$;
end

The projection model described in this section is based on the back-projection model of Dansereau et al. [18]. Although very useful for a compact lightfield representation, this model does not account for the tangential component of lateral distortion and distortion in the direction of the optical axis as in the work of Johannsen et al. [16] for focused plenoptic cameras.

5. Reconstruction

In the reconstruction problem, we want to determine the point in the object space whose rays where projected into specific points of the lightfield on the

sensor plane. Let us consider that we have a set of Z rays on the sensor plane that correspond to a given point $\mathbf{m}$ in the object space and that the intrinsic matrix $\mathbf{H}$ is known. This allows to convert the set of rays $\left\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Z\right\}$ to a set of rays $\left\{\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_Z\right\}$ in the object space.

Using the relation between a point $\mathbf{m}$ and the plane Π defined in equation (1), we obtain for the i-th ray, $\mathbf{\Phi}_i$, in the object space $x - zu_i = s_i$ and $y - zv_i = t_i$, which in matrix form corresponds to

$$\begin{bmatrix} 1 & 0 & -u_i \\ 0 & 1 & -v_i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s_i \\ t_i \end{bmatrix} \quad . \tag{10}$$

From equation (10), for each ray $\mathbf{\Phi}_i$ we obtain a set of 2 equations. The reconstruction problem has 3 unknowns to determine, hence, we need at least 2 rays to determine the corresponding point $\mathbf{m}$.

Generalizing the equation (10) for the rays $\left\{\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_Z\right\}$, and replacing those rays by the projection rays $\left\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Z\right\}$ on the sensor plane, we have

$$\begin{bmatrix} 1 & 0 & -\mathbf{h}_3\mathbf{\Phi}_1 \\ 0 & 1 & -\mathbf{h}_4\mathbf{\Phi}_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & -\mathbf{h}_3\mathbf{\Phi}_Z \\ 0 & 1 & -\mathbf{h}_4\mathbf{\Phi}_Z \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1\mathbf{\Phi}_1 \\ \mathbf{h}_2\mathbf{\Phi}_1 \\ \vdots \\ \mathbf{h}_1\mathbf{\Phi}_Z \\ \mathbf{h}_2\mathbf{\Phi}_Z \end{bmatrix} \tag{11}$$

where $\mathbf{h}_i$ corresponds to the i-th row of the intrinsic matrix $\mathbf{H}$. This is a problem that can be readily solved using a least-squares method. The equations presented in this section consider that the point $\mathbf{m}$ is defined in the camera coordinate system.

5.1. Imposing Projection Geometry Cues

The previous reconstruction methodology does not impose any prior knowledge on the correspondences $\left\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Z\right\}$ defined in the sensor plane. Hence, for a given depth of the point in the object space, the reconstruction is as good as the precision of the correspondences, maintaining all parameters of the optical system constant. Namely, due to the discretization that occurs at the image sensor, the projection rays do not define a line in the ray-space defined by the pair of coordinates (i, k) and (j, l) but a staircase (see Section 4.1, Figure 3). Therefore, the precision of the correspondences and, consequently, the reconstruction is likely to improve if we impose the projection rays in the ray-spaces to define a line. Let us call these lines in the ray-spaces as the projection geometry cues.

Let us incorporate the projection cues as a prior knowledge on the correspondences $\left\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Z\right\}$. This can be achieved by considering the point reconstruction from the lines in each of the ray-spaces (i, k) and (j, l) instead

of using the point correspondences directly. Namely, rewriting the projection equation (7) as

$$\begin{cases} \underbrace{(h_{si} + z\,h_{ui})}_{a_1}\, i + \underbrace{(h_{sk} + z\,h_{uk})}_{b_1}\, k + \underbrace{h_s + z\,h_u - x}_{c_1} = 0 \\ \underbrace{(h_{tj} + z\,h_{vj})}_{a_2}\, j + \underbrace{(h_{tl} + z\,h_{vl})}_{b_2}\, l + \underbrace{h_t + z\,h_v - y}_{c_2} = 0 \end{cases}, \tag{12}$$

we define the relation between the point in the object space and the line parameters $\boldsymbol{\theta}_{ik} = [a_1, b_1, c_1]^T$ and $\boldsymbol{\theta}_{jl} = [a_2, b_2, c_2]^T$ that define the lines in the ray-spaces (i, k) and (j, l), respectively. From these equations, one can see that, for a given point, the line parameters are fixed while the coordinates of lightfield in the sensor plane may vary. The line parameters are obtained by fitting lines to the collection of coordinate pairs (i, k) and (j, l) of the correspondences in the respective ray-space. Let us define the arrays $\boldsymbol{\Phi}_i^{ik} = [i_i, k_i, 1]^T$ and $\boldsymbol{\Phi}_i^{jl} = [j_i, l_i, 1]^T$ containing the coordinates (i, k) and (j, l) of the i-th correspondence. The line parameters can be estimated using a least-squares minimization using the Z correspondences

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{(\cdot)} &= \arg\min_{\boldsymbol{\theta}_{(\cdot)}} \sum_{i=1}^{Z} \left| \boldsymbol{\theta}_{(\cdot)}^T \boldsymbol{\Phi}_i^{(\cdot)} \right|^2 \\ &\text{s.t.} \quad \left\| \boldsymbol{\theta}_{(\cdot)} \right\|^2 = 1 \end{aligned} \tag{13}$$

where $(\cdot)$ represents either of the pair of coordinates (i, k) or (j, l) according to the ray-space that is being analyzed. These estimates for the line parameters $\hat{\boldsymbol{\theta}}_{ik}$ and $\hat{\boldsymbol{\theta}}_{jl}$ can then be used to estimate the point $\mathbf{m}$

$$\begin{bmatrix} 0 & 0 & h_{ui} & -\hat{a}_1 & 0 \\ 0 & 0 & h_{vj} & 0 & -\hat{a}_2 \\ 0 & 0 & h_{uk} & -\hat{b}_1 & 0 \\ 0 & 0 & h_{vl} & 0 & -\hat{b}_2 \\ -1 & 0 & h_u & -\hat{c}_1 & 0 \\ 0 & -1 & h_v & 0 & -\hat{c}_2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ \lambda_{ik} \\ \lambda_{jl} \end{bmatrix} = - \begin{bmatrix} h_{si} \\ h_{tj} \\ h_{sk} \\ h_{tl} \\ h_s \\ h_t \end{bmatrix} . \tag{14}$$

Remember that the line parameters are defined up to a scale factor, therefore, the scale factors λ_{ik} and λ_{jl} associated with each fitting should also be estimated to recover the correct coordinates for the point $\mathbf{m}$. This reconstruction methodology has 5 unknowns and 6 equations, which allows to obtain a solution for the point in the object space using a least squares method, for example. On the other hand, for the estimation of the line parameters, due to the constraint in equation (13), we need at least 2 correspondences to determine the 3 unknowns in each of the ray-spaces. A given correspondence contributes only with 1 equation for each of the ray-spaces. Notice that the optimization can be simplified by dividing the equations (12) by b_1 and b_2, respectively. This assumes that the singularity z_s^2 will not occur. In fact, for most of the experiments performed, this singularity occurs for points behind the camera.

For clarification purposes, we summarize the reconstruction methodology using line parameters in Algorithm 2, where we consider $\mathbf{\Phi}^{(\cdot)}$ as the collection of coordinates $(\cdot)$ of the projection rays.

Algorithm 2: Reconstruct scene point $\mathbf{m}$

Input : Projection Rays: $\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_Z\}$
Parameters: $\mathbf{H}$

Output: Scene point: $\mathbf{m} = \begin{bmatrix} x & y & z \end{bmatrix}^T$

1 Obtain $\hat{\boldsymbol{\theta}}_{ik}$ by fitting a line to $\left(\mathbf{\Phi}^i,\ \mathbf{\Phi}^k\right)$ using equation (13) ;
2 Obtain $\hat{\boldsymbol{\theta}}_{jl}$ by fitting a line to $\left(\mathbf{\Phi}^j,\ \mathbf{\Phi}^l\right)$ using equation (13) ;
3 Reconstruct $\mathbf{m}$ using equation (14)

5.2. Reconstruction Methodologies Comparison

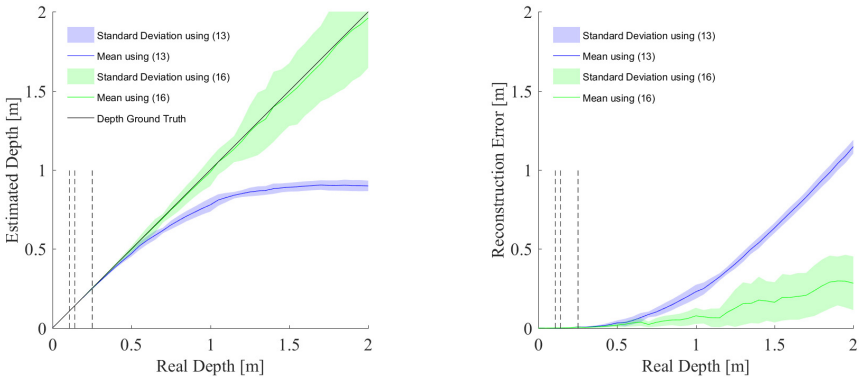
In this section, two reconstruction methodologies were proposed for point reconstruction. Let us evaluate the difference between the two methods by performing point reconstruction for points in the object space at different depths. Hence, let us consider the intrinsic matrix $\mathbf{H}_g$ provided by Dansereau et al. [18] as a result of the calibration of Dataset B. This intrinsic matrix is modified to obtain the intrinsic matrix $\mathbf{H}$ considering the geometry defined in Section 3.1 and considering the distance between the planes Π and Γ is equal to one. The entries obtained for the intrinsic matrix $\mathbf{H}$ are presented in Table 1.

h_{si}	h_{sk}	h_s	h_{tj}	h_{tl}	h_t
4.0003e-04	-9.3810e-05	1.5871e-02	3.9680e-04	-9.3704e-05	1.5867e-02
h_{ui}	h_{uk}	h_u	h_{vj}	h_{vl}	h_v
-1.5833e-03	1.9043e-03	-3.4762e-01	-1.5551e-03	1.9014e-03	-3.3817e-01

Table 1: Intrinsic matrix entries considered for evaluating the reconstruction methods.

In this experiment, the accuracy at each depth was evaluated by randomly selecting $P = 500$ points from the field of view of the plenoptic camera and computing the reconstruction error after projection and reconstruction using the two methods described. In this section, we considered an error introduced by rounding the pixels (i, j) and the microlenses (k, l) to the nearest integer. There are other alternatives to model the projection error, like for example adding noise that follows a Gaussian distribution (see Supplementary Material). The reconstruction error is defined as the distance between the reconstructed point $\hat{\mathbf{m}}_i$ and the generated point $\mathbf{m}_i$ in the object space. The mean reconstruction error r_e is defined as

$$r_e = \frac{1}{P} \sum_{i=1}^{P} \|\mathbf{m}_i - \hat{\mathbf{m}}_i\| \quad . \tag{15}$$



(a) Reconstructed Depth (b) Reconstruction Error

Figure 4: Results from reconstructing randomly generated points at depths ranging from 0.01 to 2.00 meters. The reconstructed depth is depicted in **(a)** while the reconstruction error using the (x, y, z) coordinates is depicted in **(b)**. (11) corresponds to the point reconstruction and (11) corresponds to the point reconstruction from line parameters. The dashed vertical lines, from left to right, mark the -2, -1 and 0 pixel disparities.

The depth values evaluated ranged between 0.01 and 2.00 meters. The reconstruction error and the estimated depth of these simulations are provided in Figure 4.

In Figure 4, the point reconstruction using the projection rays (blue region) start to deviate from the ground truth at 0.65 m while the reconstruction using the line parameters (green region) start to deviate from the ground truth at 1.30 m. The deviation is assumed to occur when the mean reconstruction error r_e normalized by the ground truth depth is greater than 10%. Figure 4.a shows that the mean value for the depth estimates using the projection cues are in accordance with the ground truth for the entire depth range tested. Namely, the maximum deviation from the ground truth normalized by the ground truth depth is 15.0% which is significantly lower than the 55.0% obtained for the point reconstruction applied directly to the projection rays $\mathbf{\Phi}_i$. Nonetheless, the standard deviation normalized by the ground truth depth increases significantly at 1.20 m which makes the depth estimates less reliable. Additionally, one can see that the error on the (x, y) coordinates increase more rapidly than the error on the z-coordinate with the real depth of a point.

As suggested, imposing the projection geometry cues allows to improve the depth reconstruction. More specifically, assuming the pixels (i, j) and the microlenses (k, l) are integers, the reconstruction using line parameters allows the projection ray coordinates to be real. Let us consider the depth error ε_z for a binocular stereo configuration $\varepsilon_z = \frac{z^2}{b\,f}\varepsilon_d$, where b is the baseline length, f is the focal length, z is the depth of a given point in the object space, and ε_d

corresponds to the disparity error. For a given depth of a point, maintaining all parameters constant, the depth error can only decrease by reducing the disparity error. This can be achieved by increasing the precision of the correspondences, which is achieved with the reconstruction using the line parameters. From now on, we will evaluate the depth estimation accuracy using the reconstruction from the line parameters.

6. Depth Estimation Experiments

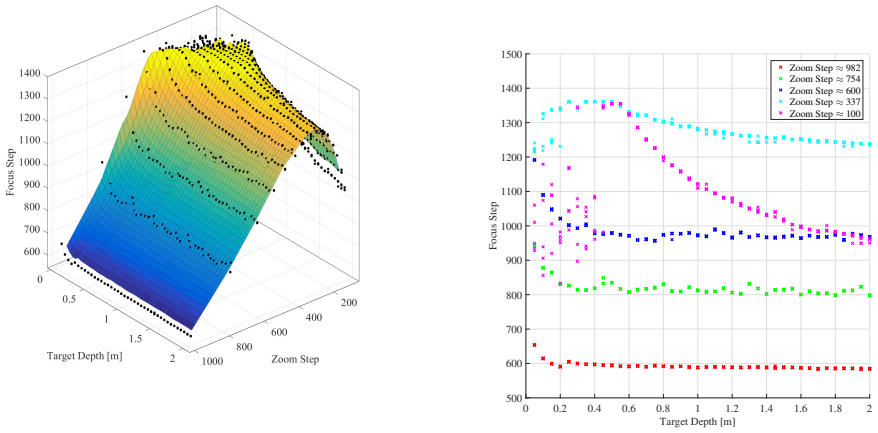
In this section, we will evaluate the reconstruction estimation accuracy for a calibrated standard plenoptic camera using the methodology defined in Section 5.1 for different zoom and focus settings.

6.1. Camera Parameters

Before proceeding to the evaluation of the reconstruction estimation accuracy, let us analyze two camera parameters that allow to determine the world focal plane. These two parameters correspond to the zoom step and focus step.

In order to analyze these parameters with the world focal plane, we compared the camera parameters with the depth of a target object in the world coordinate system. For this experiment, we acquired a set of images by placing the target object parallel to the encasing of the camera and at a regular spacing of 0.05 m from the camera. The target object depths ranged from 0.05 m to 2.00 m. The zoom number (number that appears on the interface of the camera) was changed also at a regular interval of 0.5 between 1.0 and 8.0. At each of these configurations, i.e. for a fixed target object depth and fixed zoom number, a total of 5 images were taken autofocusing on the target object. In this experiment, we are just interested in the focus step and zoom step given in the metadata of the .raw files. The results obtained are shown in Figure 5.

This figure shows that for a particular zoom step configuration, there is a depth at which the camera is not able to autofocus on the target object (the focus step does not change) and, consequently, the world focal plane does not change. This failure in focusing the target object occurs due to poor detail of the features in the image. The camera is only capable of focusing the target object, i.e., changing the world focal plane, if the zoom step is increased. Additionally, for extreme conditions of the operating range of the plenoptic camera, for example considering zoom step close to 100 and target object depths smaller than 0.4 m, one can see that the focus step changes arbitrarily among the several attempts to autofocus on the target object depth. This results in images with no sharp objects. This situation also corresponds to a failure on focusing the target object. For focusing at these target object depths, the zoom step should be decreased. This allows to conclude that the zoom also plays a role in determining the world focal plane.



(a) (b)

Figure 5: Camera autofocus given zoom step and target object depth. **(a)** The focus step is defined by autofocusing the camera on the target object. **(b)** represents the focus step with the depth of a target object for a selection of zoom steps.

6.2. Reconstruction Estimation Accuracy

Datasets. To evaluate the reconstruction estimation accuracy, we acquired seven datasets under different zoom and focus settings [1]. The zoom and focus step settings of each dataset were determined by placing a target object at a pre-determined depth of the encasing of the camera and autofocusing on this object. This allows to define a plane in focus by the main lens that is close the target object. Thus, the focus depth is assumed to be the depth of the target object. The datasets were collected using a standard plenoptic camera, the 1st generation Lytro camera. These datasets encompass images for calibration and depth range assessment. Each dataset is provided with a set of calibration plenoptic images since the camera parameters are different for each dataset. The calibration images are different from the depth images to ensure the results do not suffer from any type of overfitting effect. The calibration plenoptic images were captured using a 19×19 calibration grid of 3.18 mm cells placed at different poses and at different depths close to the target object depth bearing in mind that a minimum of 10 poses are required. On the other hand, the depth plenoptic images were captured using two different grid sizes: 19×19 grid of 6.10×6.08 mm cells and 5×7 grid of 26.50×26.38 mm cells. The grids for the depth plenoptic images were placed parallel to the encasing of the camera and at a regular spacing of 0.05 m from the camera for depth values ranging from 0.05 to 2.00 m. The two grid sizes are used for the depth plenoptic images since the

[1] www.isr.tecnico.ulisboa.pt/∼nmonteiro/datasets/plenoptic/cviu2017/

depth range evaluated is wide and it is necessary to have a reasonable number of detections to assess the depth accuracy. The smaller grid size was placed up to a maximum depth of 1.0, 1.5 and 2.0 m according to the focus depth considered 0.05, 0.50 and 1.50 m. The bigger grid size was placed considering all depth range evaluated. Table 2 summarizes the properties of the datasets acquired (for more information see Supplementary Material).

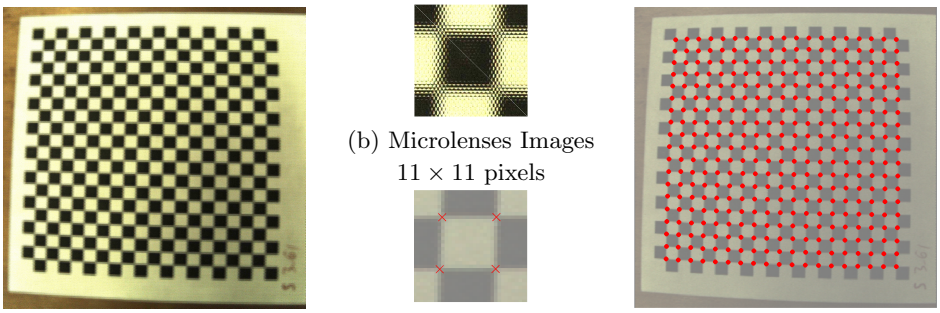
Dataset	Zoom Step	Focus Depth (m)	Calibration Depth Range (m)	Calibration Poses	Depth Poses
A	982	0.05	0.05 - 0.25	30	45
B	754	0.05	0.05 - 0.35	30	37
C	601	0.05	0.10 - 0.40	14	29
D	600	0.50	0.30 - 0.70	36	51
E	335	0.50	0.30 - 0.80	36	36
F	337	1.50	1.00 - 1.70	48	48
G	100	1.50	1.00 - 1.80	51	8

Table 2: Information regarding the zoom and focus settings of the datasets acquired for calibration and depth range assessment. The last column corresponds to the number of poses with detected features using the feature detector [34].

Calibration. Let us start by obtaining the intrinsic matrix $\mathbf{H}$ and the radial distortion parameters, the distortion center $\mathbf{r}_{uv}$ and the coefficients $k_1, \ldots, k_4$ ($M = 4$), for each dataset. The depth ranges used for the calibration procedure were defined relatively to the plane in focus by the main lens and considering the field of view of the camera. The depth range is defined relatively to the target object depth to have sharper viewpoint images which allow to detect more accurately the calibration grid points. The minimum depth value for the range was defined in order to have the full calibration grid in the viewpoint images. The number of calibration images is different among the several datasets to ensure a low ray reprojection error [18] for each dataset (see Supplementary Material). The maximum root mean-squared error for the ray reprojection error obtained during the calibration of the datasets is 0.2447 mm, which shows the accuracy of the calibration performed.

Feature Detection and Correspondences. For the evaluation of the reconstruction estimation accuracy, besides the calibration parameters we also need to know the feature points obtained for each pose of the grids captured in the plenoptic depth images. These features are the projection rays from each of the grid points captured by the camera. The projection rays are obtained by applying a feature detector [34] to each of the viewpoint images obtained from the raw plenoptic image. This is similar to the feature detection used during the calibration procedure. The major difference is that, for the plenoptic depth images, the grids may fall out of the field of view and, therefore, the number of feature points is not constant throughout all grid poses. For a lightfield with $N_i \times N_j \times N_k \times N_l$ pixels we can generate $N_i \times N_j$ viewpoint images each with

$N_k \times N_l$ pixels. The lightfield size for the standard plenoptic camera used is $11 \times 11 \times 378 \times 379$ pixels. This size is obtained after the decoding process described in [18] and removing a border of two pixels in i and j due to demosaicking and edge artifacts. Thus, among all datasets, a wide range of viewpoint images were analyzed, more precisely 58 080 viewpoint images. Although many poses were acquired for assessing the depth range of these cameras, the feature detection procedure discards many of these poses since there are no identifiable features (Table 2). Furthermore, for the depth range considered, some of the datasets only have features for a few number of depth values. The process of feature detection makes Dataset C unusable for the smaller grid size, and the Dataset G unusable for both grid sizes used.



(a) Raw Image
3280×3280 pixels

(b) Microlenses Images
11×11 pixels

(c) Detail of Viewpoint Image

(d) Viewpoint Image
381×383 pixels

Figure 6: **(a)** Debayered raw image from a standard plenoptic camera [18] with zoom **(b)** to show the effect of the microlens array. The features (k, l) obtained by the feature detector are shown in red for all calibration grid points **(d)**. The sub-pixel accuracy is depicted in **(c)**. The contrast is reduced for display.

The selection of a viewpoint image (Figure 6.d) implies the (i, j) coordinates of the projection ray, while the output from the feature detector gives us the (k, l) coordinates. The (i, j) coordinates are integers and the (k, l) coordinates are real since the feature detector has sub-pixel accuracy (Figure 6.c). The correspondences are obtained by grouping the projection rays obtained from each viewpoint image that correspond to the same grid point in the object space. Thus, each grid point has a maximum of $N_i \times N_j$ feature points. Considering the lightfield size for the plenoptic camera used, each grid point has a maximum of 121 feature points.

Camera to World Coordinate System Transformation. The methodologies described in Section 5 assume points defined in the camera coordinate system. Thus, one needs to know the rigid body transformations between the world and the camera coordinate systems defined for each of the calibrations. For each dataset, the transformation is estimated using a Procrustes analysis [35] between the estimated points in the camera coordinate system and the

ground truth points in the world coordinate system. The grids captured for each set of depth plenoptic images were only moved along the z-axis forming a parallelepiped. This allows to easily obtain the ground truth points in the world coordinate system. On the other hand, the estimated points for the grid points detected in the depth plenoptic images do not form a parallelepiped due to noise and to the reconstruction capabilities of the camera. Nonetheless, the grid points form a planar surface that is present in both coordinate systems. Hence, one can use this knowledge to remove the estimated points associated with grid depths that deviate from a planar surface. The estimated points discarded from the Procrustes analysis correspond to grid depths whose fitting error to a planar surface is above a given threshold. For a given dataset, this threshold is defined as the mean of the planar fitting errors for all grid depths in the plenoptic depth images.

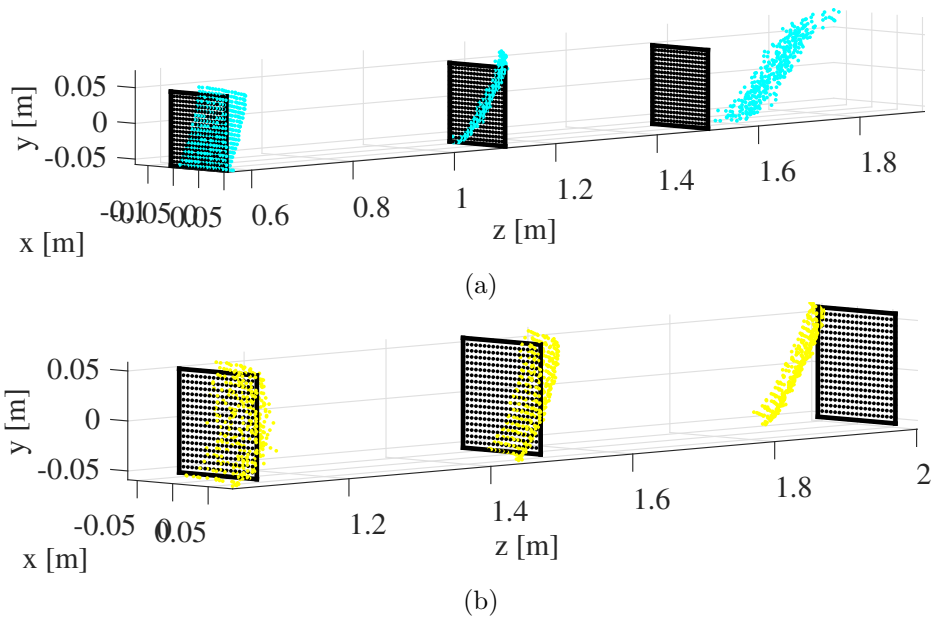


Figure 7: The estimated (cyan and yellow) and ground truth (black) grid points obtained for Datasets D and F using a smaller grid are depicted in **(a)** and **(b)**. The planar surfaces correspond to grids at the depth limits of the Datasets D and F and at an intermediate depth value (0.55 m, 1.10 m and 1.50 m for Dataset D, and 1.10 m, 1.50 m and 2.00 m for Dataset F).

In Figure 7, one can see the result of applying the estimated transformations to convert the estimated points from the camera coordinate system to the world coordinate system for three depth values of Datasets D and F using the smaller grid. Although the estimated points do not lie in a plane (reconstruction is done on a point by point basis), one can see that the estimated grids are close to planar surfaces. Additionally, comparing the grid at depth 1.5 m in each of

the datasets, one can see that the estimated points are closer to a planar surface (for Dataset D the root mean-squared error (RMSE) is 6.6 mm and for Dataset F the RMSE is 3.0 mm) and also to the corresponding ground truth grid (for Dataset D the RMSE is 0.2393 m and for Dataset F the RMSE is 0.0448 m) for Dataset F. Thus, one would expect that increasing the zoom and focus depth will originate better estimates for points farther from the camera.

Reconstruction Estimation Accuracy. Using the correspondences, the calibration parameters ($\mathbf{H}$, $\mathbf{r}_{uv}$, and $k_1, \ldots, k_4$) and the rigid body transformations, we obtain an estimate for the grid points in the world coordinate that we can use to compute the reconstruction error as defined in equation (15). The reconstruction errors and the estimated depth for the datasets are depicted in Figures 8 and 9 (additional figures can be found on the Supplementary Material). To make easier the following discussions, we also summarize in Table 3, the depth ranges identified for each of the datasets as well as the mean and standard deviation for the normalized reconstruction errors. The normalized reconstruction errors are obtained by dividing the reconstruction errors by the corresponding ground truth depths. The depth ranges are identified by determining the regions where the mean of the normalized reconstruction errors is lower or equal to 10%.

Dataset	**Depth Range (m)**	**Mean ± STD Error in Depth Range (%)**	**Mean ± STD Error (%)**
A	0.35 - 1.30	6.74 ± 5.13	16.67 ± 6.18
B	0.40 - 1.30	7.89 ± 5.96	13.72 ± 9.73
C	0.05 - 0.05	1.34 ± 5.93	25.73 ± 18.12
D	0.60 - 2.00	5.13 ± 3.20	14.01 ± 5.00
E	0.75 - 2.00	5.44 ± 3.30	8.28 ± 4.19
F	0.90 - 2.00	3.68 ± 1.78	5.90 ± 2.03
G	1.50 - 1.85	1.93 ± 0.60	1.93 ± 0.60

Table 3: Depth ranges for the datasets acquired. The depth ranges are identified as the regions whose mean for the normalized reconstruction errors is lower or equal to 10%. The mean and standard deviation (STD) for the normalized reconstruction errors within the depth ranges defined and for all ground truth depths are also depicted.

Zoom Step Analysis. In Figure 8.a-b, the datasets are grouped by constant focus depths. Namely, the figure conveys information of datasets with focus depth at 0.05 m. For this focus depth, one can see that the mean reconstruction error for points farther from the plane in focus is higher. This is also highlighted by the difference of the normalized error in the depth range and in the overall depth analyzed presented in Table 3. Additionally, one can see that the increase in zoom allows to have a lower reconstruction error for points farther from the focus depth. In Table 3, one can see that the normalized error in the whole depth

analyzed decreases while the normalized error in the depth range is maintained when the zoom is increased (excluding Dataset C due to the unusually high normalized reconstruction errors).

Focus Depth Analysis. In Figure 8.c-d, the datasets are grouped by similar zoom step. Namely, the figure conveys information of datasets with zoom step close to 336. For this zoom step, the focus depth appears to improve the reconstruction error for points at depths near the focal plane. In Table 3, this is highlighted by the change of the depth range that has a normalized reconstruction error lower or equal to 10%.

Zoom Step and Focus Depth Analysis. In Figure 9, Datasets A, D and F are depicted to highlight the reconstruction error by modifying both zoom and focus settings. In this figure, one can see that the reconstruction error decreases as the zoom step increases and the depth for which there are features detected also change. This can also be seen by the decrease on the normalized error for the whole depth analyzed and by the shift on the depth range with normalized error lower or equal to 10% in Table 3.

The lower reconstruction error with increasing zoom can be explained by considering the depth error of a binocular stereo configuration (see Section 5.2). The increase in zoom corresponds to an increase in the focal length f which leads to a decrease on the depth error, which is in accordance with the findings in this figure. On the other hand, the focus depth determines the depth at which the minimum reconstruction error occurs and, implicitly, the depth range. This can be explained looking at the ray-spaces. Namely, a point in the world focal plane corresponds to a vertical line in this space, which leads to a smaller error due to the discretization that occurs at the image sensor (staircase effect). As the point moves away from the world focal plane, the line starts to deviate from this vertical line and the discretization error increases leading to an increase on the reconstruction error. Notice that the reconstruction method presented in Section 5.1 reduces but does not eliminate the reconstruction error associated with discretization.

The reconstruction results presented in this section are obtained considering the radial distortion parameters. The mean difference of the estimated points normalized by the ground truth depth by not including the radial distortion parameters is less than 1.6% for all datasets analyzed. This difference does not change significantly the results presented in Table 3 (see Supplementary Material). Thus, we consider that the radial distortion does not play an important role on the reconstruction estimation accuracy.

The results presented in this section show that the standard plenoptic cameras has a reconstruction estimation accuracy that varies with the zoom and focus settings of the camera. The zoom is a determinant factor on increasing the reconstruction accuracy of these cameras, while the focus depth (as a combination of zoom and focus steps, see Figure 5) plays a role on shifting the depth range. The depth range analyzed from 0.05 m to 2.00 m can be reconstructed

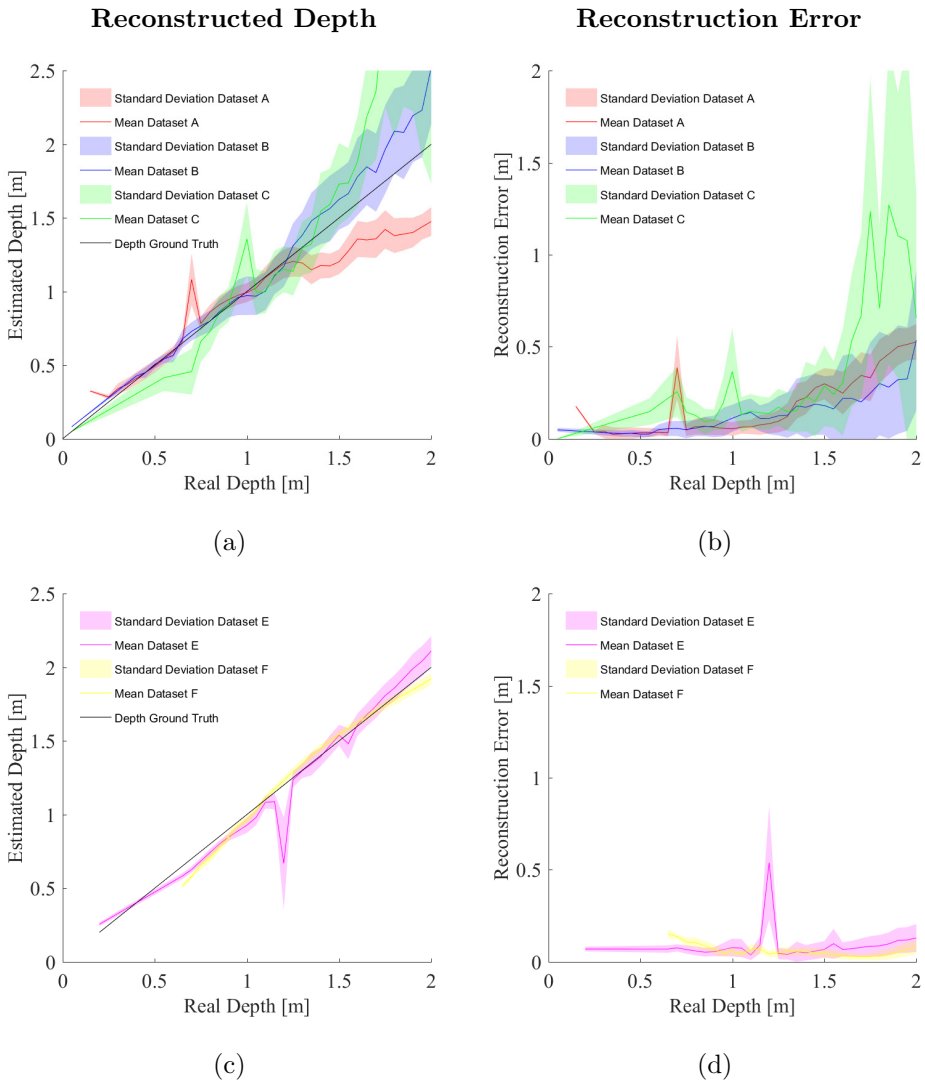


Figure 8: Reconstruction estimation accuracy with zoom step **(first row)** and with focus depth **(second row)**. The **first column** depicts the reconstructed depth while the **second column** depicts the reconstruction error for the estimated points. The **first row** groups the datasets with focus depth at 0.05 m (Datasets A, B and C) and the **second row** groups the datasets with zoom step close to 336 (Datasets E and F).

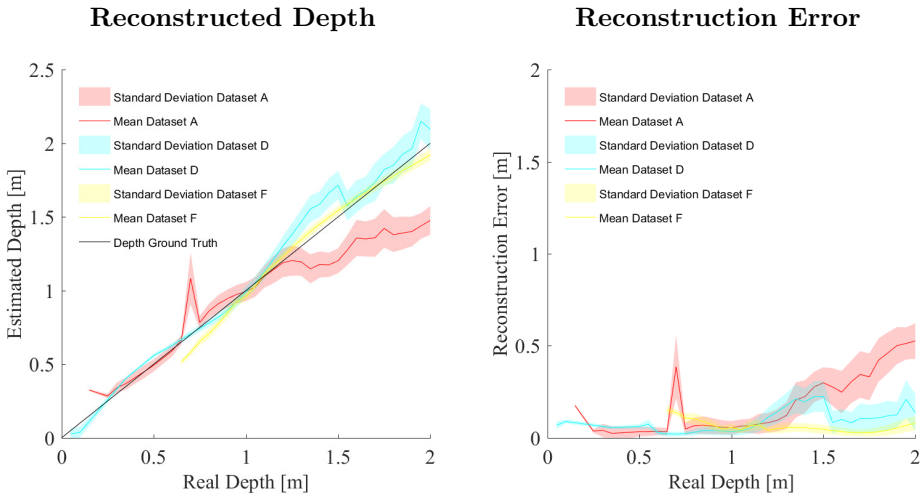


Figure 9: Reconstruction estimation accuracy with zoom step and focus depth. The **first column** depicts the reconstructed depth while the **second column** depicts the reconstruction error for the estimated points obtained for datasets with different zoom and focus settings.

with accuracy by choosing correctly the zoom and focus settings of the camera.

7. Conclusions

In this work, we extended the camera model of Dansereau et al. [18] to formalize a forward projection model and improved a reconstruction methodology based on multi-view stereo by imposing geometry cues on the ray-spaces (i, k) and (j, l). These reconstruction methods assume that we have a calibrated standard plenoptic camera.

The improved reconstruction methodology was used to evaluate the depth estimation accuracy under certain zoom and focus settings. This method was applied to seven datasets acquired with different zoom and focus settings and with objects placed at depths ranging from 0.05 to 2.00 meters. The depth accuracy was evaluated through the reconstruction of grid points captured on these datasets (using features on viewpoint images). The findings presented suggest that these cameras are capable of reconstructing points in the depth range analyzed by appropriately choosing the zoom and focus settings. Namely, the zoom increase allows to lower the reconstruction error while the focus depth determines the depth range of the camera. This is the first study, to the best of our knowledge, that studies the depth capabilities of a standard plenoptic camera for different zoom and focus settings.

Feature detection is an important component of the reconstruction methodology. As shown experimentally, the inclusion of projection geometric cues provided already a significant improvement of the depth reconstruction results. Further improvements on feature detection, on the projection model (by including other types of distortion), or on the reconstruction methodology, can be found which could allow extending the range of reconstructible depths for a specific error bound. The reconstruction analysis can be further complemented by defining a theoretical error and studying the changes introduced by different zoom and focus step settings on the intrinsics matrix **H**.

Acknowledgements

Funding: This work was supported by the Portuguese Foundation for Science and Technology (FCT) project [grant numbers UID/EEA/50009/2013, and PD/BD/105778/2014], and by the CMU-Portugal Project Augmented Human Assistance (AHA) [grant number CMUP-ERI/HCI/0046/2013].

References

[1] S. Ogata, J. Ishida, T. Sasano, Optical sensor array in an artificial compound eye, Optical Engineering 33 (11) (1994) 3649–3655.

[2] D. Floreano, R. Pericet-Camara, S. Viollet, F. Ruffier, A. Brückner, R. Leitel, W. Buss, M. Menouni, F. Expert, R. Juston, et al., Miniature curved artificial compound eyes, Proceedings of the National Academy of Sciences 110 (23) (2013) 9267–9272.

[3] E. R. Dowski Jr, G. E. Johnson, Wavefront coding: a modern method of achieving high-performance and/or low-cost imaging systems, in: SPIE's International Symposium on Optical Science, Engineering, and Instrumentation, International Society for Optics and Photonics, 1999, pp. 137–145.

[4] E. H. Adelson, J. Y. A. Wang, Single lens stereo with a plenoptic camera, IEEE Transactions on Pattern Analysis & Machine Intelligence (2) (1992) 99–106.

[5] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, Light field photography with a hand-held plenoptic camera, Computer Science Technical Report CSTR 2 (11) (2005) 1–11.

[6] R. Ng, Digital light field photography, Ph.D. thesis, stanford university (2006).

[7] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, C. Intwala, Spatio-angular resolution tradeoffs in integral photography., Rendering Techniques 2006 (2006) 263–272.

[8] A. Lumsdaine, T. Georgiev, The focused plenoptic camera, in: Computational Photography (ICCP), 2009 IEEE International Conference on, IEEE, 2009, pp. 1–8.

[9] C. Perwass, L. Wietzke, Single lens 3d-camera with extended depth-of-field, in: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 2012, pp. 829108–829108.

[10] D. Dansereau, L. Bruton, Gradient-based depth estimation from 4d light fields, in: Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on, Vol. 3, IEEE, 2004, pp. III–549.

[11] M. Diebold, B. Goldluecke, Epipolar plane image refocusing for improved depth estimation and occlusion handling.

[12] S. Wanner, B. Goldluecke, Variational light field analysis for disparity estimation and super-resolution, Pattern Analysis and Machine Intelligence, IEEE Transactions on 36 (3) (2014) 606–619.

[13] J. Lüke, F. Rosa, J. Marichal-Hernández, J. Sanluı, C. Domı, J. Rodrı, et al., Depth from light fields analyzing 4d local structure, Journal of Display Technology 11 (11) (2015) 900–907.

[14] N. B. Monteiro, J. P. Barreto, J. Gaspar, Dense lightfield disparity estimation using total variation regularization, in: International Conference Image Analysis and Recognition, Springer, 2016, pp. 462–469.

[15] M. W. Tao, S. Hadap, J. Malik, R. Ramamoorthi, Depth from combining defocus and correspondence using light-field cameras, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 673–680.

[16] O. Johannsen, C. Heinze, B. Goldluecke, C. Perwaß, On the calibration of focused plenoptic cameras., in: Time-of-Flight and Depth Imaging, Springer, 2013, pp. 302–317.

[17] N. Zeller, F. Quint, U. Stilla, Calibration and accuracy analysis of a focused plenoptic camera, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2 (3) (2014) 205.

[18] D. G. Dansereau, O. Pizarro, S. B. Williams, Decoding, calibration and rectification for lenselet-based plenoptic cameras, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 1027–1034.

[19] O. Johannsen, A. Sulc, B. Goldluecke, On linear structure from motion for light field cameras, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 720–728.

[20] C. Hahne, A. Aggoun, S. Haxha, V. Velisavljevic, J. C. J. Fernández, Light field geometry of a standard plenoptic camera, Optics express 22 (22) (2014) 26659–26673.

[21] C. Hahne, A. Aggoun, V. Velisavljevic, S. Fiebig, M. Pesch, Refocusing distance of a standard plenoptic camera, Optics Express 24 (19) (2016) 21521–21540.

[22] R. C. Bolles, H. H. Baker, D. H. Marimont, Epipolar-plane image analysis: An approach to determining structure from motion, International Journal of Computer Vision 1 (1) (1987) 7–55.

[23] M. V. Afonso, J. M. Bioucas-Dias, M. A. Figueiredo, An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems, Image Processing, IEEE Transactions on 20 (3) (2011) 681–695.

[24] J. Yu, L. McMillan, S. Gortler, Scam light field rendering, in: Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on, IEEE, 2002, pp. 137–144.

[25] C. Chen, H. Lin, Z. Yu, S. Bing Kang, J. Yu, Light field stereo matching using bilateral statistics of surface cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1518–1525.

[26] O. Johannsen, K. Honauer, B. Goldluecke, A. Alperovich, F. Battisti, Y. Bok, M. Brizzi, M. Carli, G. Choe, M. Diebold, et al., A taxonomy and evaluation of dense light field depth estimation algorithms, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2017, pp. 82–99.

[27] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, R. Ramamoorthi, Shape estimation from shading, defocus, and correspondence using light-field angular coherence, IEEE transactions on pattern analysis and machine intelligence 39 (3) (2017) 546–560.

[28] M. Levoy, P. Hanrahan, Light field rendering, in: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM, 1996, pp. 31–42.

[29] S. J. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen, The lumigraph, in: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM, 1996, pp. 43–54.

[30] E. H. Adelson, J. R. Bergen, The plenoptic function and the elements of early vision, Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

[31] D. G. Dansereau, I. Mahon, O. Pizarro, S. B. Williams, Plenoptic flow: Closed-form visual odometry for light field cameras, in: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, IEEE, 2011, pp. 4455–4462.

[32] M. D. Grossberg, S. K. Nayar, The raxel imaging model and ray-based calibration, International Journal of Computer Vision 61 (2) (2005) 119–137.

[33] S. Wanner, B. Goldluecke, Globally consistent depth labeling of 4d light fields, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 41–48.

[34] A. Kassir, T. Peynot, Reliable automatic camera-laser calibration, in: Proceedings of the 2010 Australasian Conference on Robotics & Automation, ARAA, 2010.

[35] D. G. Kendall, A survey of the statistical theory of shape, Statistical Science (1989) 87–99.