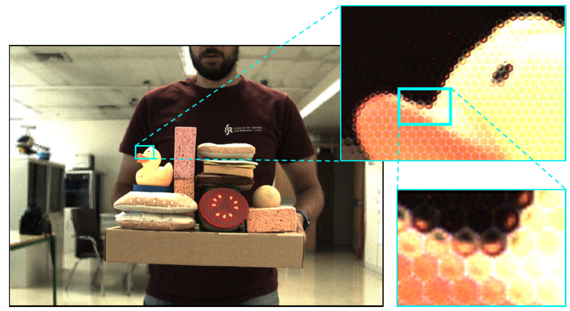
Draft submitted to IEEE Transactions on Circuits and Systems for Video Technology. Accepted November 2019.
Online: https://ieeexplore.ieee.org/document/8906167

Standard Plenoptic Cameras Mapping to Camera Arrays and Calibration based on DLT

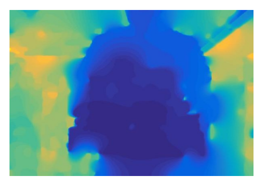
Nuno Barroso Monteiro, Joao P. Barreto, José António Gaspar, *Member, IEEE*

***Abstract*—First prototypes of standard plenoptic cameras (SPCs) were based on arrays of pinhole cameras. Despite the array nature, viewpoint pinhole arrays are not intrinsically provided by current SPC calibration tools. In this work, we start by detailing the mapping between the SPC model and a camera array of viewpoints. Then, the mapping is used to propose a calibration procedure for the SPC based on a grid of corners. Calibration involves two steps, first a linear solution and then a nonlinear optimization minimizing the ray re-projection error. The proposed calibration methodology compares favourably with state of the art calibrations and the linear solution proposed for the initial stage of the calibration outperforms the state of the art.**

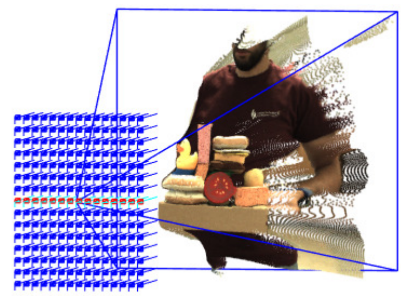
***Index Terms*—Standard Plenoptic Camera, Viewpoint Camera Array, DLT Calibration.**



(a) SPC raw image and zoom of microlenses



(b) Reconstructed depth map



(c) 3D reconstruction and camera array (centers spaced $50\times$)

Fig. 1: Lightfield scene reconstruction and camera array. **(a)** Image captured on the sensor of a SPC with detail of the MIs formed in the sensor. **(b)** depicts the depth map obtained using [1]. **(c)** Viewpoint camera array obtained by calibration where the spacing among projection centers has been scaled 50 times to be perceptible on the 3D plot.

I. Introduction

IN a pinhole camera, different light rays reflected by a point in the object space are captured in a single pixel location in the image space. In a plenoptic camera, these different light ray directions are captured at different pixels, therefore creating a lightfield [2], [3]. Lightfields open new possibilities like single image depth estimation [4], [5] or refocusing [6]. Relevance and interest of these applications motivated the appearance of several types of lenslet based plenoptic cameras as the standard plenoptic camera (SPC) [7] or the focused plenoptic camera (FPC) [8]. Comprehensive introduction and review of the major lightfield concepts and capabilities can be found in overview articles as [9], [10].

In this work, we focus on the SPC which consists of a main lens, one single high definition imaging sensor, and a microlens array. The SPC geometry generates unfocused microlens images (MIs) (Figure 1.a) by placing the main lens focal plane on the microlens array plane [11].

The geometry model most used for SPCs is the one proposed by Dansereau *et al.* [12]. This model maps rays in the image space indexed by pixels and microlenses indices to rays in the object space defined in metric units. The concept of viewpoint image (VI) defined by Ng *et al.* [6], obtained by selecting the same pixel for each microlens, allows to conveniently view the SPC as a camera array (Figure 1.c). Camera arrays help explaining the geometry of viewpoint cameras. However, the projection model for the viewpoint cameras is still to be fully formalized and there is no connection established with the camera model proposed by Dansereau *et al.* [12].

In this work, we build from the model of Dansereau *et al.* [12] and derive the mapping between a SPC and the viewpoint camera array. The viewpoint camera array representation is used to define a new calibration procedure. The accuracy of the mapping described is evaluated by a corner based calibration for commercially available SPCs. The code and datasets used are provided [1].

Contributions. The contributions of this work are two-fold: (i) formal definition of the projection model for a viewpoint

This work was supported by the Portuguese Foundation for Science and Technology (FCT) projects (grant numbers UID/EEA/50009/2019, PD/BD/105778/2014, and PINFRA/22084/2016) and the European Commission project ORIENT (ERC/2016/693400). J. P. Barreto thanks FCT and COMPETE2020 for generous funding under grant PTDC/EEIAUT/3024/2014.

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

J. P. Barreto is with the Institute for Systems and Robotics, University of Coimbra, Portugal, jpbar@isr.uc.pt

[1] www.isr.tecnico.ulisboa.pt/∼nmonteiro/articles/plenoptic/tcsvt2019/

camera, and mapping between the models of a SPC proposed by Dansereau *et al*. [12] and an array of viewpoint pinhole cameras, and (ii) definition of a linear solution for a SPC capable of estimating all parameters of the camera model based on the viewpoint camera array representation. This work can also be seen as an entry point to plenoptic cameras for researchers and developers acquainted with the pinhole camera model.

In terms of structure, we present in Section II a review of the camera array mappings and calibration procedures for SPCs. In Section III, we introduce the SPC model removing the redundancies with the extrinsic parameters. The mapping from the SPC model to the viewpoint camera array representation is described in Section IV. The proposed calibration procedure is presented in Section V with an emphasis on the linear solution obtained based on the viewpoint camera array representation. The results of applying the calibration proposed to calibrate comercially available SPCs are reported in Section VI and the major conclusions are presented in Section VII.

Notation: non-italic letters correspond to functions, italic letters correspond to scalars, lower case bold letters correspond to vectors, and upper case bold letters correspond to matrices. Vectors represented in homogeneous coordinates are denoted by $\tilde{(\cdot)}$.

II. Related Work

The camera models proposed for SPCs [12], [13] consider the microlenses as pinholes and the main lens as a thin lens. Recent works define a projection matrix associated with each microlens of the microlens array. Namely, Bok *et al*. [13] describes the microlens array for a SPC using six parameters and the knowledge of the corresponding microlenses centers in the raw image. The microlenses centers are not assumed to be regularly spaced as in Dansereau *et al*. [12]. Zeller *et al*. [14] also describes the microlens camera array but for a FPC with the purpose of enabling visual odometry directly from the MIs.

Bok *et al*. [13] performs calibration based on line features extracted from the MIs on the raw image. This method is not adequate to calibrate the SPC when the calibration grid is placed near the world focal plane of the main lens because of the difficulty of detecting features on the unfocused MIs. On the other hand, the calibration procedure proposed by Dansereau *et al*. [12] is capable of calibrating the SPC even on this situation.

In Dansereau *et al*. [12], the lightfield in the image space defined in pixels (i, j) and microlenses (k, l) indices is mapped to the lightfield in the object space defined by a position (s, t) and a direction (u, v) in metric units (Figure 2). This mapping considers a 5×5 matrix with ten free intrinsic parameters that is obtained by propagating the rays from the sensor to the object space using ray transfer matrices. Nonetheless, there is not provided a direct connection with a projection matrix for either the microlens or viewpoint cameras. That connection, detailed in this work, allows adapting methods from the mainstream computer vision to plenoptic cameras.

The 5×5 matrix is used to calibrate a virtual SPC using corner points on VI as features. The linear solution for the calibration procedure described in [12] is based on estimating an homography for each viewpoint camera and pose of the calibration pattern. This solution estimates eight from the ten free intrinsic parameters, being the remaining two parameters estimated later in the nonlinear optimization.

The calibration with VIs requires a pre-processing step, denoted as decoding [12], to transform the 2D raw image (Figure 1.a) into a 4D lightfield. There are several approaches for the decoding process like the ones presented in [12], [15], [16]. In this work, we focus solely on the calibration of a SPC. The decoding originates a virtual SPC that assumes a lightfield that is obtained considering that the microlenses define a rectangular tiling instead of the actual hexagonal tiling (Figure 1.a).

The closer connection of the mapping proposed by Dansereau *et al*. [12] to a pinhole projection matrix is the one provided by Marto *et al*. [1] regarding the representation of a camera array composed of identical co-planar cameras. However, the mapping proposed in [1] does not explain the zero disparity in the epipolar plane images (EPIs) for points in the main lens world focal plane (box B in Figure 3.b) [17].

There are few works referring to the geometry of the viewpoint cameras. In Hahne *et al*. [18], the location of the viewpoints projection centers is defined using the same ray propagation strategy from sensor to object space as in Dansereau *et al*. [12]. Nonetheless, the complete viewpoint projection matrix is not defined and no association with a SPC model is made since the optical settings of the main lens and microlens array are assumed to be known. This work is tailored to aid in the design of new plenoptic cameras. In Bok *et al*. [13], a first attempt is made to define the projection matrix for the viewpoints. However, the intrinsic matrix is assumed to be common among all viewpoint cameras and is defined based on the diameter (in pixels) of a MI and the knowledge of the parameters used to describe the microlens array rather than using a geometrically approach. Furthermore, the geometry proposed does not allow to explain the zero disparity for points in the world focal plane of the main lens.

In this work, we consider the pinhole viewpoint camera constraint to represent the mapping introduced in [12] using eight free intrinsic parameters. This is accomplished by shifting the rays parameterization plane along the optical axis of the camera [19] to the plane containing the viewpoint projection centers. Additionally, we provide the mapping between the virtual SPC and the viewpoint camera array that is consistent with the zero disparity in the EPI for points in the world focal plane of the main lens. The viewpoint camera array representation is used to define a calibration for the SPC based on corner features from VIs. The linear solution proposed starts from the estimation of a single generalized homography for all viewpoints per pose of the pattern, and extending techniques from pinhole camera calibration recover the eight free intrinsic parameters of the camera model. Generalization is obtained considering a camera array composed of different co-planar cameras with parameterized principal point shift and baseline among viewpoints.

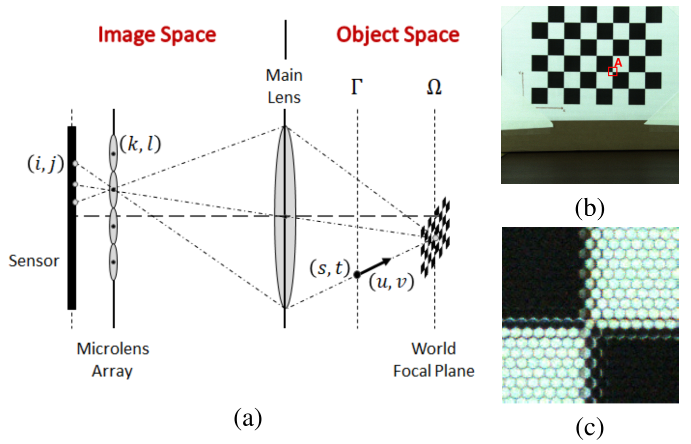


Fig. 2: Geometry of a SPC whose main lens focal plane corresponds to plane Ω **(a)**. The lightfield in the image space is parameterized using pixels (i, j) and microlenses (k, l) indices while the lightfield in the object space is parameterized using a point (s, t) defined on the parameterization plane Γ and a direction (u, v). **(b)** shows the raw image of a calibration grid placed on the mains lens world focal plane and **(c)** exhibits the details of the microlenses in red box A.

III. Standard Plenoptic Camera

A SPC can be represented by a 5×5 matrix $\mathbf{H}$ [12] which maps back-projection rays on sensor coordinates to rays in object (metric) space coordinates. In formal terms, $\mathbf{H}$ is a mapping of rays $\tilde{\mathbf{\Phi}} = [i, j, k, l, 1]^T$ in the image space to rays $\tilde{\mathbf{\Psi}} = [s, t, u, v, 1]^T$ in the object space:

$$\tilde{\mathbf{\Psi}} = \mathbf{H}\ \tilde{\mathbf{\Phi}} \tag{1}$$

where rays $\mathbf{\Phi}$ are parameterized using pixels (i, j) and microlenses (k, l) indices and rays $\mathbf{\Psi}$ are parameterized using a position (s, t) on a plane Γ and a direction (u, v) defined in metric units [17] (Figure 2.a). The mapping defined by Dansereau *et al.* [12] has 12 non-zero entries, however choosing the plane Γ to coincide with the plane containing the viewpoint projection centers (Supp. Material B) and removing the redundancies with the translational components of the extrinsic parameters (Supp. Material C) allows to define the mapping $\mathbf{H}$ with 8 non-zero entries

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

In order to help establishing the relationship between the SPC and the pinhole camera model, in the following we denominate $\mathbf{H}$ as the lightfield intrinsics matrix (LFIM). We note that LFIM is a simplified term, as $\mathbf{H}$ effectively contains intrinsic parameters information, however, it also contains baseline information, as detailed in Section IV. Conventional extrinsic parameters, as found in pinhole camera models, defining a world coordinate system, are in fact not contained in $\mathbf{H}$.

One ray $\mathbf{\Psi} = [s, t, u, v]^T$ can be represented as one parametric 3D line [20], namely $[x, y, z]^T = [s, t, 0]^T + \lambda[u, v, 1]^T$ for $\lambda \in \mathbb{R}$. Therefore, the LFIM matrix (2) allows to define the relationship between an arbitrary point $[x, y, z]^T$ in the object space and the ray $\mathbf{\Phi}$ in the image space [17] as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{H}^{st}_{ij} \begin{bmatrix} i \\ j \end{bmatrix} + z \left(\mathbf{H}^{uv}_{ij} \begin{bmatrix} i \\ j \end{bmatrix} + \mathbf{H}^{uv}_{kl} \begin{bmatrix} k \\ l \end{bmatrix} + \mathbf{h}_{uv} \right) \tag{3}$$

where the LFIM is partitioned in three 2×2 sub-matrices and one 2×1 vector $\mathbf{h}_{uv} = [h_u, h_v]^T$. The sub-matrices follow the notation $\mathbf{H}^{(\cdot)}_{(\cdot)}$ where the subscript selects the columns and the superscript selects the lines, *i.e.* for example, $\mathbf{H}^{st}_{ij}$ selects the first two columns, denoted by ij, and the first two lines, denoted by st. Equation (3) shows that given one ray in image coordinates, the LFIM $\mathbf{H}$ allows defining a back-projection ray in the object space or, equivalently, one 3D point at a specific depth z.

IV. Viewpoint Camera Array

In this section, we represent a SPC as a camera array of viewpoints. The array representation is mapped from the SPC model defined by Dansereau *et al.* [12] for SPCs, namely, from the LFIM (2).

Let the projection matrix $\mathbf{P}^{ij}$, parameterized by the coordinates $(i, j) \in \mathbb{Z}^2$, represent the SPC as an array

$$\mathbf{P}^{ij} = \mathbf{K}^{ij} \begin{bmatrix} \mathbf{I}_{3\times 3} & \mathbf{t}^{ij} \end{bmatrix} {}^{c}\mathbf{T}_w \tag{4}$$

where $\mathbf{K}^{ij}$ denotes the intrinsic matrix, $\mathbf{I}_{3\times3}$ is a 3×3 identity matrix, $\mathbf{t}^{ij}$ is the projection center and ${}^{c}\mathbf{T}_w = \begin{bmatrix} {}^{c}\mathbf{R}_w & {}^{c}\mathbf{t}_w \\ \mathbf{0}_{1\times3} & 1 \end{bmatrix}$ defines the rigid body transformation between the world and camera coordinate systems with rotation ${}^{c}\mathbf{R}_w \in SO(3)$ and translation ${}^{c}\mathbf{t}_w \in \mathbb{R}^3$, and $\mathbf{0}_{1\times3}$ is the 1×3 null matrix.

Note that while ${}^{c}\mathbf{T}_w$ defines one coordinate system for all viewpoints, the intrinsic matrix and the projection center are different for each viewpoint camera (i, j). In the following, let the camera model for the viewpoint cameras (4) take into account that the principal point and the projection center are different for each viewpoint while the scale factor remains the same:

$$\mathbf{K}^{ij} = \begin{bmatrix} k_u & 0 & u_0 + i\ \Delta u_0 \\ 0 & k_v & v_0 + j\ \Delta v_0 \\ 0 & 0 & 1 \end{bmatrix} \text{and } \mathbf{t}^{ij} = \begin{bmatrix} i\ \Delta x_0 \\ j\ \Delta y_0 \\ 0 \end{bmatrix} \tag{5}$$

where the scalars k_u and k_v denote focal lengths and conversion from metric units to pixels (denominated as scale factors in the remainder of the paper). The vector $[u_0, v_0]^T$ defines the principal point for viewpoint $(i, j) = (0, 0)$, and the vectors $[\Delta u_0, \Delta v_0]^T$ and $[\Delta x_0, \Delta y_0, 0]^T$ denote principal point shift and baseline between consecutive viewpoint cameras, respectively.

A. Mapping from LFIM to Viewpoint Projection Matrices

Considering that the rays of one viewpoint camera converge to a unique point (s,t) (Supp. Material B), one may set constant the values (i,j) and solve equation (3) relatively to (k,l). This gives an equation of a viewpoint pixel (k,l) imaging the 3D point (x,y,z) that can be rewritten as a pinhole model, equations (4) and (5), with the intrinsic matrix and the projection center defined as

$$\mathbf{K}^{ij} = \begin{bmatrix} \frac{1}{h_{uk}} & 0 & -\frac{h_u}{h_{uk}} - i\ \frac{h_{ui}}{h_{uk}} \\ 0 & \frac{1}{h_{vl}} & -\frac{h_v}{h_{vl}} - j\ \frac{h_{vj}}{h_{vl}} \\ 0 & 0 & 1 \end{bmatrix} \text{and}\, \mathbf{t}^{ij} = \begin{bmatrix} -i\ h_{si} \\ -j\ h_{tj} \\ 0 \end{bmatrix} . \quad (6)$$

This allows to obtain the mappings to the representations in (5). Namely, comparing (6) with (5), we identify a common component $[u_0, v_0]^T = -\left[h_u/h_{uk}, h_v/h_{vl}\right]^T$ and a differential (shift) component $[\Delta u_0, \Delta v_0]^T = -\left[h_{ui}/h_{uk}, h_{vj}/h_{vl}\right]^T$ on the principal point. The scale factors are defined as $k_u = 1/h_{uk}$ and $k_v = 1/h_{vl}$, and the baseline is defined as $[\Delta x_0, \Delta y_0, 0]^T = -\left[h_{si}, h_{tj}, 0\right]^T$.

An example of the pinhole model parameters for a viewpoint camera array obtained from a calibrated Lytro Illum camera can be found in Table V. This array is configured for a focused depth of about 1.09 meters and describes 15×15 cameras, $i \in \{1, \dots, 15\}$, equal for j, whose VIs have 625×433 (k,l) pixels.

B. Properties of Viewpoint Projection Matrices

Considering equation (3), one can obtain the EPI geometry that relates the depth of a point with the disparity on the VIs $\left[\frac{\Delta k}{\Delta i}, \frac{\Delta l}{\Delta j}\right]^T$

$$\frac{\Delta k}{\Delta i} = -\frac{h_{si}}{h_{uk}}\frac{1}{z} - \frac{h_{ui}}{h_{uk}} \quad \text{and} \quad \frac{\Delta l}{\Delta j} = -\frac{h_{tj}}{h_{vl}}\frac{1}{z} - \frac{h_{vj}}{h_{vl}} . \quad (7)$$

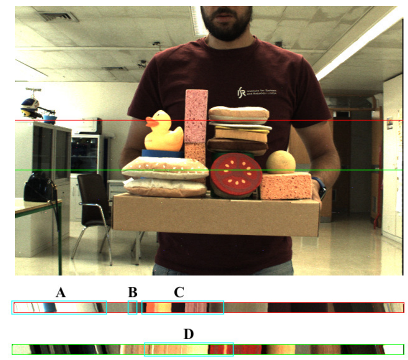
The mapping (6) allows to rewrite the EPI geometry defined in equation (7) as

$$\frac{\Delta k}{\Delta i} = k_u \frac{\Delta x_0}{z} + \Delta u_0 \quad \text{and} \quad \frac{\Delta l}{\Delta j} = k_v \frac{\Delta y_0}{z} + \Delta v_0 . \quad (8)$$

The EPI geometry shows that despite the parallel optical axis, the zero disparity plane, also known as the optical focal plane [6] of the SPC main lens is at a finite depth due to the principal point shift (box B in Figure 3.b). Considering the geometry of the camera in Figure 2.a, the zero disparity plane corresponds to the plane Ω with $z_\Omega = -k_u \frac{\Delta x_0}{\Delta u_0} = -k_v \frac{\Delta y_0}{\Delta v_0}$.

Contrarily, if we consider the principal point shift equal to zero, *i.e*. cameras with same principal point and therefore same intrinsic matrix $\mathbf{K}^{ij}$, one recovers the EPI geometry defined in [21] that defines points at infinity as the points of zero disparity [17]. Looking at the EPIs obtained from a lightfield in Figure 3, one can see that the lines corresponding to different points in the object space have a range of positive and negative slopes. Namely, objects in the background (box A) have a negative slope while objects in the foreground (box C and D) have a positive slope. The disparity zero, in this scene, corresponds to the position of the person holding the objects (box B).

Notice also that the field of view is similar for all viewpoint cameras. Scene regions observed by the different viewpoint cameras change slightly for depths other than the zero disparity plane depth z_Ω (Figure 4.d). This is a consequence of the array of projection centers and array of principal points modeling viewpoint cameras. For the zero disparity plane depth $z_\Omega = -\frac{h_{si}}{h_{ui}} = -\frac{h_{tj}}{h_{vj}}$, the influence of the different projection centers is cancelled by the principal point shift and the scene region observed is the same for all viewpoint cameras (Figure 4.c).



(a)
(b)
(c)

Fig. 3: The viewpoint cameras identified in red in Figure 1.c are used to obtain EPIs from the lightfield at rows 185 (red) **(b)** and 265 (green) **(c)** on the central viewpoint **(a)**.

V. Standard Plenoptic Camera Calibration

The calibration proposed considers the corners of a planar calibration grid of known dimensions as features. In the following, we assume that the corners in the world coordinate system have been matched with the imaged corners. Let us consider that we have a 4D lightfield obtained from the raw image (Figure 2.a) after the decoding process [12], [22]. An imaged corner is defined by a ray $\mathbf{\Phi} = [i,j,k,l]^T$ in the image space. The (k,l) coordinates correspond to the pixel coordinates of the detected corners on the VIs. The (i,j) coordinates correspond to the viewpoint coordinates.

A. Linear Initialization

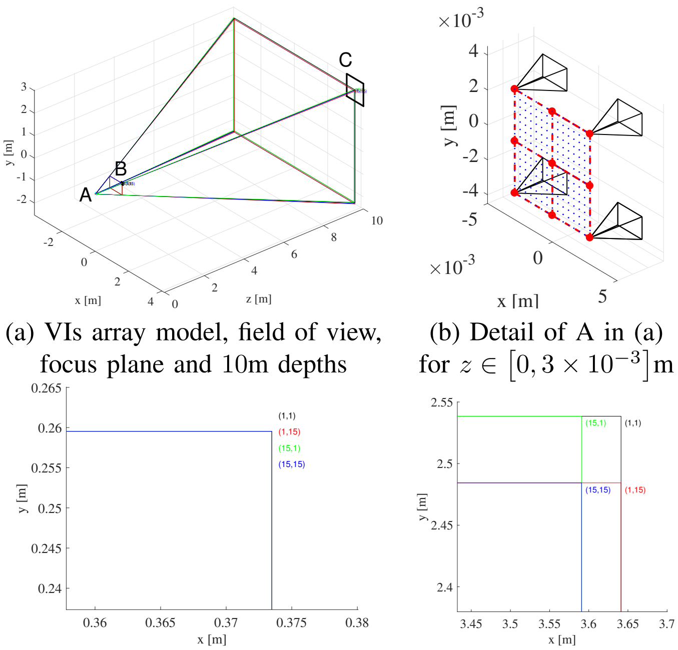
In this section, we will consider the mapping in Section IV to define a linear solution for the LFIM $\mathbf{H}$ associated with a plenoptic camera and the extrinsic parameters for each pose of the calibration grid.

Homography Estimation. Considering the viewpoint projection matrix (4), a point $\mathbf{m} = [x,y,z]^T$ in the object space is projected to a point in the image plane $\mathbf{q}$ by

$$\tilde{\mathbf{q}} \sim \mathbf{P}^{ij}\tilde{\mathbf{m}} = \mathbf{K}^{ij}\left[{}^{c}\mathbf{R}_w \quad {}^{c}\mathbf{t}_w + \mathbf{t}^{ij}\right]\tilde{\mathbf{m}} \quad (9)$$

where the symbol $\sim$ denotes equal up to a scale factor. The coplanar grid points allow to define a world coordinate system such that the z-coordinate is zero. In this context, denoting $\tilde{\mathbf{m}} = [x,y,1]^T$, one can redefine the projection (9) as $\tilde{\mathbf{q}} \sim \mathbf{H}^{ij}\tilde{\mathbf{m}}$ where

$$\mathbf{H}^{ij} = \mathbf{K}^{ij}\left[\mathbf{r}_1 \quad \mathbf{r}_2 \quad {}^{c}\mathbf{t}_w + \mathbf{t}^{ij}\right] \quad (10)$$



(a) VIs array model, field of view, focus plane and 10m depths

(b) Detail of A in (a) for $z \in \left[0, 3 \times 10^{-3}\right]$m

(c) Detail of B in (a)

(d) Detail of C in (a)

Fig. 4: Field of view of a Lytro Illum camera analyzed from the VIs array model. **(a)** back-projection pyramids of the four corner VIs, $(i,j) = \left\{(1,1), (1,15), (15,1), (15,15)\right\}$, where A represents the array of projection centers, B is at the focus plane at depth z_Ω, and C is at depth $z = 10$m. **(b)** zoom of A in (a), other VIs projection centers shown by red lines and blue dots. **(c)** zoom of black rectangle B in (a) showing the region observed at z_Ω is the same for all VIs. **(d)** zoom of black rectangle C in (a) shows slight differences of regions observed by the different viewpoint cameras.

is the parametric homography matrix for the viewpoint camera (i,j), and ${}^{c}\mathbf{R}_w = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$. This matrix can be estimated from the point correspondences $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$ using a direct linear transformation (DLT) [23]. Each point correspondence originates 2 linearly independent equations. The homography matrix has 9 entries to estimate but is defined only up to scale. Thus, $\mathbf{H}^{ij}$ has 8 degrees of freedom needing at least 4 point correspondences to estimate its entries [24]. Assuming a plenoptic camera with N pixels within each microlens and considering an independent estimation of each of the viewpoint cameras' homography matrices, one has $8N$ unknowns to estimate.

A plenoptic camera introduces restrictions on the viewpoint camera array that allows to decrease the number of unknowns to estimate. Namely, the homography matrix $\mathbf{H}^{ij}$ change among viewpoints as a result of the principal point shift and baseline in (6). Let us consider that $\mathbf{H}^{ij}$ can be defined from the homography matrix $\mathbf{H}^0$ associated with the viewpoint coordinates $(i,j) = (0,0)$ and the homography viewpoint change matrix $\mathbf{A}^{ij}$ by

$$\mathbf{H}^{ij} = \underbrace{\begin{bmatrix} h^0_{11} & h^0_{12} & h^0_{13} \\ h^0_{21} & h^0_{22} & h^0_{23} \\ h^0_{31} & h^0_{32} & h^0_{33} \end{bmatrix}}_{\mathbf{H}^0} + \begin{bmatrix} i & 0 & 0 \\ 0 & j & 0 \\ 0 & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 0 \end{bmatrix}}_{\mathbf{A}^{ij}}. \tag{11}$$

Considering the homography projection of a calibration grid corner $\tilde{\mathbf{m}} = [x, y, 1]^T$ in the object space to the image point $\tilde{\mathbf{q}}$ for the viewpoint camera (i,j), applying the cross product by $\tilde{\mathbf{q}}$ on each side of the projection equation leads to $[\tilde{\mathbf{q}}]_\times \mathbf{H}^{ij}\tilde{\mathbf{m}} = \mathbf{0}_{3\times1}$, where $\left[(\cdot)\right]_\times$ is a skew-symmetric matrix that applies the cross product. Using the properties of the Kronecker product [25] and solving for each of the unknown parameters, one obtains

$$\left(\tilde{\mathbf{m}}^T \otimes [\tilde{\mathbf{q}}]_\times\right) \mathbf{T} \begin{bmatrix} \mathbf{h}^0 \\ \mathbf{a}^{ij} \end{bmatrix} = \mathbf{0}_{3\times1} \tag{12}$$

where

$$\mathbf{T} = \left[\mathbf{I}_{9\times9} \; \begin{matrix} i & 0 & 0 & 0 & 0 & 0 \\ 0 & j & 0 & 0 & 0 & 0 \\ \multicolumn{6}{c}{\mathbf{0}_{1\times6}} \\ 0 & 0 & i & 0 & 0 & 0 \\ 0 & 0 & 0 & j & 0 & 0 \\ \multicolumn{6}{c}{\mathbf{0}_{1\times6}} \\ 0 & 0 & 0 & 0 & i & 0 \\ 0 & 0 & 0 & 0 & 0 & j \\ \multicolumn{6}{c}{\mathbf{0}_{1\times6}} \end{matrix} \right], \tag{13}$$

and $\mathbf{h}^0$ and $\mathbf{a}^{ij}$ correspond to vectorizations of the matrix $\mathbf{H}^0$ and $\mathbf{A}^{ij}$ by stacking their columns and removing the zero entries, respectively. The solution $\left[\mathbf{h}^0, \mathbf{a}^{ij}\right]^T$ for the parametric homography matrix can be estimated using singular value decomposition (SVD) (Supp. Material D).

The restrictions introduced by a plenoptic camera allow to represent the parametric homography matrix (11) using 15 parameters. According to equation (12), each point correspondence $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$ originates three equations with only two being linearly independent. On the other hand, each point in the object space originates N image points, one for each viewpoint camera, assuming that the point is observed in all viewpoint cameras. These pairs provide $2N$ equations that, theoretically, are enough to estimate the parametric homography matrix, assuming that $N \geq 8$. Nonetheless, the restrictions on the viewpoint camera array also originate restrictions on the projections of a point in the object space. Namely, the ray in the image space $\mathbf{\Phi}^{ij} = [i, j, k, l]^T$ associated with an arbitrary viewpoint (i,j) can be described from the ray coordinates $\mathbf{\Phi}^0 = [0, 0, k_0, l_0]^T$ associated with the viewpoint $(i,j) = (0,0)$ by $\mathbf{\Phi}^{ij} = \mathbf{\Phi}^0 + [i, j, i\beta, j\beta]^T$, where β corresponds to the disparity of the point defined on the VIs. This reduces the number of linearly independent equations originated by a point in the object space to 4. Thus, one needs at least 4 non-collinear points to obtain the entries of the homography matrix $\mathbf{H}^{ij}$.

Intrinsic and Extrinsic Estimation. The structure of the homography matrix (10) in conjunction with the orthogonality and identity of the column vectors of ${}^{c}\mathbf{R}_w$ allow to define constraints on the intrinsic parameters as ${\mathbf{h}_1}^T\mathbf{B}^{ij}\mathbf{h}_2 = 0$ and ${\mathbf{h}_1}^T\mathbf{B}^{ij}\mathbf{h}_1 - {\mathbf{h}_2}^T\mathbf{B}^{ij}\mathbf{h}_2 = 0$ [26] where $\mathbf{h}_n$ refers to the n-th column vector of $\mathbf{H}^{ij}$, and the symmetric matrix that describes the image of the absolute conic is defined as $\mathbf{B}^{ij} = {\mathbf{K}^{ij}}^{-T}{\mathbf{K}^{ij}}^{-1}$ [26], [27]. These constraints can be used to obtain the intrinsic parameters independently for each of the viewpoint cameras [26]. Alternatively, one can use the knowledge of the intrinsic matrix defined in Section IV-A to perform the estimation of a parametric representation of the absolute conic $\mathbf{B}^{ij}$ for a viewpoint camera (i, j) using a minimal number of parameters.

The intrinsic matrix $\mathbf{K}^{ij}$ differs on the principal point for each viewpoint leading to different images of the absolute conic. The principal points change regularly between consecutive viewpoints by $\left[-\frac{h_{ui}}{h_{uk}}, -\frac{h_{vj}}{h_{vl}}\right]^T$ which can be used to constraint the parametric representation of $\mathbf{B}^{ij}$. Namely, considering (6), $\mathbf{B}^{ij}$ can be defined as

$$\mathbf{B}^{ij} = \mathbf{B}^0 + i\ \mathbf{C}^i + j\ \mathbf{D}^j + i^2\ \mathbf{E}^i + j^2\ \mathbf{F}^j \tag{14}$$

with

$$\mathbf{B}^0 = \begin{bmatrix} h_{uk}^2 & 0 & h_u h_{uk} \\ 0 & h_{vl}^2 & h_v h_{vl} \\ h_u h_{uk} & h_v h_{vl} & 1 + h_u^2 + h_v^2 \end{bmatrix}, \tag{15}$$

$$\mathbf{C}^i = \begin{bmatrix} 0 & 0 & h_{ui}h_{uk} \\ 0 & 0 & 0 \\ h_{ui}h_{uk} & 0 & 2h_u h_{ui} \end{bmatrix}, \mathbf{E}^i = \begin{bmatrix} & \mathbf{0}_{2\times 3} & \\ 0 & 0 & h_{ui}^2 \end{bmatrix}, \tag{16}$$

$$\mathbf{D}^j = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & h_{vj}h_{vl} \\ 0 & h_{vj}h_{vl} & 2h_v h_{vj} \end{bmatrix}, \text{ and } \mathbf{F}^j = \begin{bmatrix} & \mathbf{0}_{2\times 3} & \\ 0 & 0 & h_{vj}^2 \end{bmatrix}. \tag{17}$$

This allows to define a representation for $\mathbf{B}^{ij}$ using 11 distinct non-zero entries $\mathbf{b}^{ij} = [\, b_{11}, b_{13}, b_{22}, b_{23}, b_{33}, c_{13}, c_{33}, d_{23}, d_{33}, e_{33}, f_{33}\,]^T$ where $(\cdot)_{mn}$ represents the entry in row m and column n of the matrix $(\cdot)$. Considering these parameters, the intrinsic parameters constraints can be redefined as

$$\begin{bmatrix} h_{11}h_{12} & {h_{11}}^2 - {h_{12}}^2 \\ h_{11}h_{32} + h_{12}h_{31} & 2\left(h_{11}h_{31} - h_{12}h_{32}\right) \\ h_{21}h_{22} & {h_{21}}^2 - {h_{22}}^2 \\ h_{21}h_{32} + h_{22}h_{31} & 2\left(h_{21}h_{31} - h_{22}h_{32}\right) \\ h_{31}h_{32} & {h_{31}}^2 - {h_{32}}^2 \\ i\left(h_{11}h_{32} + h_{12}h_{31}\right) & 2i\left(h_{11}h_{31} - h_{12}h_{32}\right) \\ i\left(h_{31}h_{32}\right) & i\left({h_{31}}^2 - {h_{32}}^2\right) \\ j\left(h_{21}h_{32} + h_{22}h_{31}\right) & 2j\left(h_{21}h_{31} - h_{22}h_{32}\right) \\ j\left(h_{31}h_{32}\right) & j\left({h_{31}}^2 - {h_{32}}^2\right) \\ i^2\left(h_{31}h_{32}\right) & i^2\left({h_{31}}^2 - {h_{32}}^2\right) \\ j^2\left(h_{31}h_{32}\right) & j^2\left({h_{31}}^2 - {h_{32}}^2\right) \end{bmatrix}^T \mathbf{b}^{ij} = \mathbf{0}_{2\times 1}. \tag{18}$$

Normally, each homography generates 2 equations for determining the matrix of the absolute conic image [26]. The parametric representation (11), representing an arbitrary viewpoint (i, j), generates 6 equations. Nonetheless, only 2 equations are independent regarding the entries of $\mathbf{B}^0$, so one needs to acquire at least 3 calibration grid poses to estimate $\mathbf{b}^{ij}$ defined up to a scale factor.

The intrinsic matrix parameters can be recovered from $\mathbf{B}^{ij}$. More specifically, rewriting the intrinsic matrix $\mathbf{K}^{ij}$ (6) as

$$\mathbf{K}^{ij} = \underbrace{\begin{bmatrix} \frac{1}{h_{uk}} & 0 & -\frac{h_u}{h_{uk}} \\ 0 & \frac{1}{h_{vl}} & -\frac{h_v}{h_{vl}} \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K}^0} + \begin{bmatrix} i & 0 & 0 \\ 0 & j & 0 \\ 0 & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} & -\frac{h_{ui}}{h_{uk}} \\ \mathbf{0}_{3\times 2} & -\frac{h_{vj}}{h_{vl}} \\ & 0 \end{bmatrix}}_{\mathbf{G}^{ij}}, \tag{19}$$

one can define $\mathbf{B}^0 = {\mathbf{K}^0}^{-T}{\mathbf{K}^0}^{-1}$. This allows to estimate the entries of $\mathbf{K}^0$ using the Cholesky decomposition of $\mathbf{B}^0$ and correcting the scale factor considering $k_{33}^0 = 1$. The principal point shift can be estimated considering $\frac{h_{ui}}{h_{uk}} = \frac{e_{33}^i}{c_{13}}$ and $\frac{h_{vj}}{h_{vl}} = \frac{f_{33}^j}{d_{23}}$.

The extrinsic parameters can be estimated once the intrinsic matrix $\mathbf{K}^{ij}$ is known. From (10), the rotation matrix ${}^c\mathbf{R}_w = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$ is recovered considering

$$\mathbf{r}_1 = \lambda {\mathbf{K}^{ij}}^{-1}\mathbf{h}_1\ ,\ \mathbf{r}_2 = \lambda {\mathbf{K}^{ij}}^{-1}\mathbf{h}_2\ ,\ \text{and } \mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2\ . \tag{20}$$

with $\lambda = 1/\left\|{\mathbf{K}^{ij}}^{-1}\mathbf{h}_1\right\| = 1/\left\|{\mathbf{K}^{ij}}^{-1}\mathbf{h}_2\right\|$. The translation and projection center $\mathbf{t}^{ij}$ are recovered solving the following system of equations

$$\lambda\mathbf{h}_3 = \begin{bmatrix} \mathbf{K}^{ij} & -i\mathbf{k}_1 & -j\mathbf{k}_2 \end{bmatrix} \begin{bmatrix} {}^c\mathbf{t}_w \\ h_{si} \\ h_{tj} \end{bmatrix} \tag{21}$$

where $\mathbf{k}_n$ corresponds to the n-th column of the parametric intrinsic matrix $\mathbf{K}^{ij}$.

B. Nonlinear Optimization

In this section, the linear solution is refined and radial distortion is considered on the coordinates (u, v). Namely, the undistorted rays in the object space $\mathbf{\Psi}^u = [s, t, u^u, v^u]^T$ are defined from distorted rays in the object space $\mathbf{\Psi} = [s, t, u, v]^T$ by

$$\begin{bmatrix} u^u \\ v^u \end{bmatrix} = \left(1 + k_1\ r^2 + k_2\ r^4 + k_3\ r^6\right) \begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} + \begin{bmatrix} b_u \\ b_v \end{bmatrix} \tag{22}$$

where $\dot{u} = u^u - b_u$, $\dot{v} = v^u - b_v$, $r^2 = u^2 + v^2$, and $\mathbf{d} = (k_1, k_2, k_3, b_u, b_v)$ defines the distortion vector. In the distortion vector, k_1, k_2 and k_3 are the radial distortion correction coefficients while the vector $[b_u, b_v]^T$ defines the distortion center. In the nonlinear optimization, we minimize the ray re-reprojection error. This optimization refines the intrinsic parameters $\mathbf{H}$, the extrinsic parameters $\mathbf{R}_m$ (parameterized by Rodrigues formula [28]) and $\mathbf{t}_m$, $m = 1, \ldots, M$ where M is the number of poses, and the distortion vector $\mathbf{d}$:

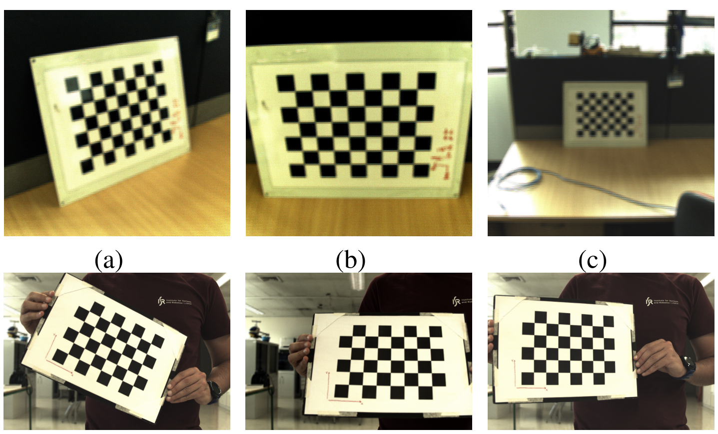
$$\underset{\mathbf{H}, \mathbf{R}_m, \mathbf{t}_m, \mathbf{d}}{\arg\min} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \Lambda\left(\eta_n\left(\mathbf{H}, \mathbf{d}\right), \mathbf{R}_m\ \mathbf{m}_n + \mathbf{t}_m\right) \tag{23}$$

where N_m corresponds to the number of corners detected on a pose m, $\Lambda(\cdot)$ defines the point-to-ray distance [12], η defines the direction coordinates (u^u, v^u) after mapping the ray in the image space $\mathbf{\Phi}_n$ associated with the corner n to the ray in object space (equation (1)) and followed by distortion rectification (equation (22)). $\mathbf{m}_n$ defines the 3D corner point in the world coordinate system. The nonlinear optimization is solved using the trust-region-reflective algorithm [29], where a sparsity pattern for the Jacobian matrix is provided. The number of parameters over which we optimize is 8 for the intrinsic parameters, 5 for the lens distortion parameters, and $6M$ for the extrinsic parameters.

VI. Experimental Results

The calibration methodology proposed in Section V is assessed in this section using calibration datasets acquired with commercially available SPCs: the 1^{st} generation Lytro camera and the most recent Lytro Illum.

Plenoptic cameras acquire images that have higher storage requirements than conventional cameras. Namely, the 1^{st} generation Lytro has a raw image with 3280×3280 pixels (Figure 5.a-c) that allows to define 9×9 VIs with a resolution of 383×381 pixels after the decoding process described in [12], [22]. On the other hand, the Lytro Illum has a higher spatial and angular resolution in consequence of the higher number of microlenses in the sensor and the higher number of pixels within each microlens. More specifically, the raw image has 7728×5368 pixels (Figure 5.d-e) that allows to define 15×15 VIs with a resolution of 625×433 pixels after the decoding process described in [12], [22].



(a) (b) (c)

(d) (e) (f)

Fig. 5: Calibration data (raw images) from 1^{st} generation Lytro camera [12] **(a)-(c)** and Lytro Illum **(d)-(e)**.

A. 1^{st} Generation Lytro State of the Art Comparison

In this section, we compare the results of the calibration procedure proposed in Section V with the calibrations proposed by Dansereau *et al*. [12] (denoted as *Dans13*) and Bok *et al*. [13] (denoted as *Bok17*). The calibration procedures are applied to publicly available calibration datasets [12] that were obtained using a 1^{st} generation Lytro camera. For this comparison, we considered the root mean square (RMS) of the re-projection error, the ray re-projection error (*i.e.* distance between ray and 3D point as defined in [12]), and the reconstruction error, for three stages of the calibration process: the initial linear solution (Section V-A), the nonlinear refinement, with and without distortion estimation (Section V-B). Three errors are used in this comparison since the re-projection error is the usual error while evaluating pinhole camera calibration procedures but, in plenoptic cameras, the error normally used is the ray re-projection error [12], [13]. In addition, the reconstruction error is used to assess the quality of the reconstruction at the different stages of the calibration. The errors are summarized in Tables I, II, and III. Notice that the values from Bok *et al*. [13] are retrieved directly from their paper.

Comparing the results of the calibration proposed with the ones obtained using *Dans13* [12], one can see that the major

RMS Re-Projection Error (pix)		Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
Initial	*Dans13* [12]	(10) 1.678 (5) 1.673	(18) 1.687 (5) 1.695	(12) 1.687 (5) 1.671	(10) 1.748 (5) 1.714	(17) 4.290 (5) 4.700
	Bok17* [13]	-	-	-	-	-
	Ours	(10) 0.838 (5) **0.797**	(18) **0.856** (5) 1.035	(12) **0.950** (5) 0.953	(10) 0.965 (5) **0.790**	(17) 0.840 (5) **0.627**
Optimized	*Dans13* [12]	(10) 0.435 (5) 0.372	(18) 0.406 (5) 0.429	(12) 0.402 (5) **0.392**	(10) 0.404 (5) 0.461	(17) 0.218 (5) 0.185
	Bok17* [13]	-	-	-	-	-
	Ours	(10) 0.427 (5) **0.366**	(18) **0.405** (5) 0.435	(12) 0.420 (5) **0.392**	(10) **0.389** (5) 0.489	(17) 0.219 (5) **0.177**
Optimized (with Distortion)	*Dans13* [12]	(10) 0.226 (5) 0.221	(18) 0.191 (5) 0.240	(12) 0.161 (5) 0.164	(10) 0.150 (5) 0.163	(17) 0.190 (5) 0.153
	Bok17* [13]	(5) 0.374	(9) 0.259	-	-	(14) 0.274
	Ours	(10) 0.226 (5) **0.211**	(18) **0.179** (5) 0.194	(12) **0.156** (5) 0.159	(10) **0.145** (5) 0.163	(17) 0.134 (5) **0.127**

TABLE I: RMS re-projection error in pixels for three stages of the calibration procedure: initial linear solution, and nonlinear refinement with and without distortion estimation. The number of poses M considered for the calibration is denoted as (M). The symbol * indicates that the values reported are retrieved directly from the corresponding paper.

RMS Ray Re-Projection Error (mm)		Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
Initial	*Dans13** [12]	(10) 3.200	(18) 5.060	(12) 8.630	(10) 5.920	(17) 13.800
	Dans13 [12]	(10) 0.577 (5) 0.627	(18) 0.603 (5) 0.570	(12) 1.036 (5) 0.974	(10) 1.231 (5) 1.081	(17) 8.900 (5) 11.970
	Bok17* [13]	-	-	-	-	-
	Ours	(10) **0.307** (5) 0.314	(18) **0.341** (5) 0.353	(12) 0.609 (5) **0.593**	(10) 0.640 (5) **0.478**	(17) **1.657** (5) 1.709
Optimized	*Dans13** [12]	(10) 0.146	(18) 0.148	(12) 0.255	(10) **0.247**	(17) **0.471**
	Dans13 [12]	(10) 0.154 (5) 0.145	(18) 0.147 (5) **0.139**	(12) 0.260 (5) **0.245**	(10) 0.260 (5) 0.268	(17) 0.485 (5) 0.546
	Bok17* [13]	-	-	-	-	-
	Ours	(10) 0.151 (5) **0.143**	(18) 0.143 (5) **0.139**	(12) 0.271 (5) 0.247	(10) 0.251 (5) 0.277	(17) 0.489 (5) 0.532
Optimized (with Distortion)	*Dans13** [12]	(10) **0.084**	(18) **0.063**	(12) 0.106	(10) **0.105**	(17) **0.363**
	Dans13 [12]	(10) 0.085 (5) 0.086	(18) 0.066 (5) 0.069	(12) 0.104 (5) **0.102**	(10) 0.116 (5) 0.117	(17) 0.390 (5) 0.456
	Bok17* [13]	(5) 0.108	(9) 0.071 (5) 0.072	-	-	(14) 0.492 (5) 0.454
	Ours	(10) 0.085 (5) 0.085	(18) 0.066 (5) 0.066	(12) 0.103 (5) 0.103	(10) 0.114 (5) 0.116	(17) 0.393 (5) 0.457

TABLE II: RMS ray re-projection error in mm for three stages of the calibration procedure: initial linear solution, and nonlinear refinement with and without distortion estimation. As in Table I, (M) denotes M poses, and * indicates values retrieved from related work.

RMS Reconstruction Error (mm)		Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
Initial	*Dans13* [12]	(10) 2100.536 (5) 3139.904	(18) 325.215 (5) 203.736	(12) 1293.985 (5) 1397.874	(10) 844.038 (5) 517.783	(17) 2702.292 (5) 3370.725
	Ours	(10) **3.039** (5) 3.904	(18) **6.212** (5) 8.023	(12) 14.899 (5) **12.558**	(10) **20.751** (5) 25.316	(17) **79.681** (5) 102.281
Optimized	*Dans13* [12]	(10) **3.370** (5) 3.627	(18) 4.367 (5) **3.112**	(12) 10.174 (5) 10.607	(10) 15.050 (5) 12.401	(17) **123.728** (5) 253.959
	Ours	(10) 3.747 (5) 3.682	(18) 4.516 (5) 3.927	(12) 10.229 (5) **8.277**	(10) 15.168 (5) **12.216**	(17) 142.231 (5) 187.750
Optimized (with Distortion)	*Dans13* [12]	(10) 4.408 (5) 4.283	(18) 4.652 (5) **4.415**	(12) 9.995 (5) 8.007	(10) 15.425 (5) 13.051	(17) **135.851** (5) 179.697
	Ours	(10) 4.443 (5) **4.256**	(18) 4.706 (5) **4.415**	(12) 9.976 (5) **7.932**	(10) 15.553 (5) **12.700**	(17) 138.968 (5) 183.037

TABLE III: RMS reconstruction error in mm for three stages of the calibration procedure: initial linear solution, and nonlinear refinement with and without distortion estimation. As in Tables I and II, (M) denotes M poses.

differences occur at the linear solution. For a subset of 5 poses of these datasets, in the linear solution stage, one can see that the re-projection error of *Dans13* [12] is at least 1.63 times higher, the ray re-projection error is at least 1.61 times higher, and the reconstruction error is at least 20.45 times higher. These differences between the two calibration methods are even greater when we consider the complete datasets. This confirms that the proposed method for the initial linear solution outperforms the state of the art.

Comparing with Bok *et al*. [13], the proposed calibration obtains smaller re-projection and ray re-projection errors using the complete datasets. Namely, the re-projection error is 1.44 smaller, and the ray re-projection error is 1.25 times smaller. Only Dataset B presents a similar performance to the calibration proposed. Considering a subset of 5 poses, the ray re-projection errors obtained for Bok *et al*. [13] are similar with the ones of the calibration proposed with the exception of Dataset A that exhibits an error 1.26 times higher.

The results obtained show that the entries h_{sk} and h_{tl} can be set to zero without degrading the performance of the calibration procedure. The position (s,t) of the ray can be represented using only the viewpoint coordinates (i,j) if the parameterization plane corresponds to the plane containing the viewpoint projection centers (Supp. Material C). This allows to represent the rays with a minimal number of sub-camera apertures, and the LFIM with a minimal number of parameters.

Concluding, the characterization of the viewpoint camera array for the 1^{st} generation Lytro camera (Dataset B [12]) is presented in Table V. These parameters are obtained from the LFIM estimated at the final stage of the calibration procedure. The camera array is characterized by a unitary baseline length $\left\|\mathbf{t}^{ij}\right\| = \sqrt{\Delta x_0^2 + \Delta y_0^2}$ of 0.37 mm. Considering the 9×9 viewpoint cameras, the maximum baseline length that can be defined is 2.97 mm. The non-zero principal point shift shows that the principal point is different for each viewpoint camera. This gives a zero disparity 3D plane, *i.e*. the plane in focus Ω, positioned approximately at 0.29 m for Dataset B [12].

B. Calibration Precision with Number of Poses

As in the calibration of conventional pinhole cameras, the redundancy and accuracy of calibration data is a key factor for attenuating the effect of calibration data noise into the calibration precision. Dansereau *et al*. [12] considered the influence of different sizes of calibration patterns while Bok *et al*. [13] considered the influence of two different number of poses: 5 poses or the full calibration dataset.

The Lytro Illum camera is more recent than the 1^{st} generation Lytro camera, and its specifications indicate improvement in almost all technical aspects. However, the previous analysis were only performed with the 1^{st} generation Lytro camera. Thus, in this section, we want to assess the influence of the different sizes of the calibration patterns and different number of poses. For this purpose, we acquired new calibration datasets with a Lytro Illum camera using two calibration grids with different sizes: 8×6 grid of 211×159 mm with approximately 26.5 mm cells (denoted as Illum-1), and 20×20 grid of 121.5×122 mm with approximately 6.1 mm cells (denoted as Illum-2). Each dataset acquired is composed of 66 fully observable poses of the calibration pattern. Care was taken to avoid changing the focal settings of the camera.

The higher number of poses acquired allow to define several subsets for calibration which allow a statistical analysis of the results. Therefore, in order to evaluate the precision of the calibration, we repeated 20 times the calibration procedure. Each calibration involves $k = 2, \ldots, 20$ pattern poses, randomly selected from the full calibration dataset. The calibration procedure proposed in Section V is compared with the methodology [12] (denoted as *Dans13*).

The mean and standard deviation obtained for the re-projection error, the ray re-projection error, and the reconstruction error with the number of poses for Dataset Illum-1 and Illum-2 are depicted in Figure 6. This figure shows that the errors are similar for both calibration methods after nonlinear refinement. However, for the initial linear solution, the calibration proposed obtains smaller errors using 3 or more calibration pattern poses.

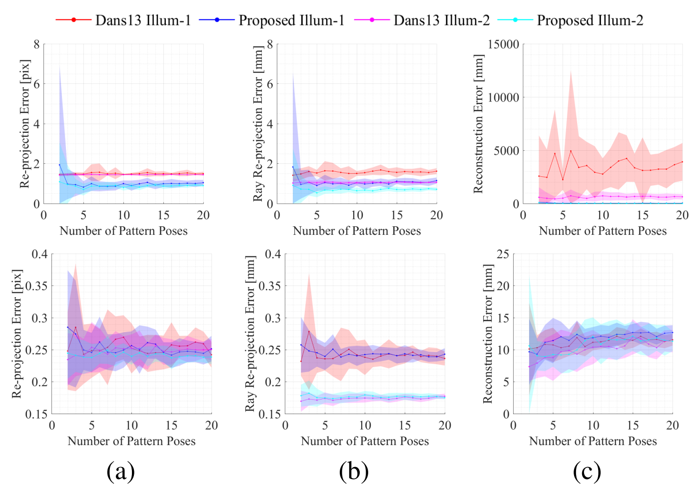


Fig. 6: RMS errors obtained using the calibration proposed (in blue and cyan for Dataset Illum-1 and Illum-2, respectively) and the calibration *Dans13* [12] (in red and magenta for Dataset Illum-1 and Illum-2, respectively): re-projection error **(a)**, ray re-projection error **(b)**, and reconstruction error **(c)**. The first row depicts the errors obtained for the initial linear solution and the second row depicts the errors obtained for the nonlinear refinement with distortion estimation.

The calibration proposed (Section V) is applied to a set of 10 randomly sampled poses and on a reduced set of 5 poses to evaluate the quality of the calibration. For comparison purposes, these sets of poses are also calibrated using the calibration described by Dansereau *et al*. [12]. For this comparison, we considered the RMS of the re-projection error, the ray re-projection error (as defined in [12]), and the reconstruction error, for three stages of the calibration process: the initial linear solution, the nonlinear refinement, with and without distortion estimation. The errors are summarized in Table IV.

The re-projection, ray re-projection and reconstruction errors are similar after the refinement of the initial linear solution for both calibration methods. Also, the lower number of

RMS Re-Projection Error (pix)		Illum-1 10 poses	Illum-1 5 poses	Illum-2 10 poses	Illum-2 5 poses
Initial	Dans13 [12]	1.463	1.501	1.480	1.485
	Ours	1.659	1.400	**0.922**	1.249
Optimized	Dans13 [12]	**0.320**	0.428	0.418	0.429
	Ours	0.332	0.429	0.421	0.446
Optimized (with Distortion)	Dans13 [12]	**0.235**	0.288	0.293	0.288
	Ours	0.249	0.284	0.263	0.270
RMS Ray Re-Projection Error (mm)		**Illum-1 10 poses**	**Illum-1 5 poses**	**Illum-2 10 poses**	**Illum-2 5 poses**
Initial	Dans13 [12]	1.698	1.516	0.965	0.891
	Ours	1.813	1.623	**0.617**	0.776
Optimized	Dans13 [12]	0.322	0.342	**0.245**	0.255
	Ours	0.334	0.347	0.247	0.261
Optimized (with Distortion)	Dans13 [12]	0.241	0.239	0.168	0.173
	Ours	0.254	0.243	**0.166**	0.172
RMS Reconstruction Error (mm)		**Illum-1 10 poses**	**Illum-1 5 poses**	**Illum-2 10 poses**	**Illum-2 5 poses**
Initial	Dans13 [12]	3483.898	1553.119	304.433	199.785
	Ours	37.594	18.717	**7.560**	9.626
Optimized	Dans13 [12]	13.625	9.046	8.126	**6.496**
	Ours	13.747	10.563	8.242	6.914
Optimized (with Distortion)	Dans13 [12]	10.680	10.255	7.070	**5.968**
	Ours	11.939	10.526	6.850	6.250

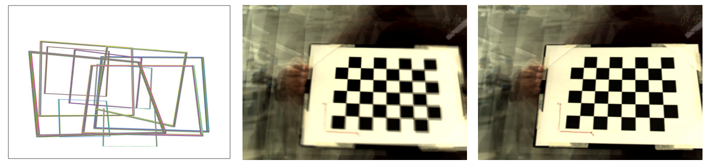
TABLE IV: RMS errors for three stages of the calibration procedure: initial linear solution, and nonlinear refinement with and without distortion estimation.

poses does not appear to change the errors significantly after the nonlinear optimization. The accuracy of the calibration proposed can be seen from the maximum errors obtained at the final stage of the calibration: the re-projection error has sub-pixel accuracy (below 0.29 pixels), the ray re-projection error is below 0.26 mm, and the reconstruction error is below 12 mm.

For the initial linear solution, the re-projection and ray re-projection errors are similar for the Dataset Illum-1. However, for the Dataset Illum-2, one can see that these errors are smaller for the calibration proposed. Additionally, the reconstruction error is considerably higher for the calibration proposed by Dansereau *et al*. [12] regardless of the dataset considered. More specifically, the reconstruction error is at least 20 times higher than the one obtained using the calibration proposed.

The major difference between *Dans13* [12] and the proposed method corresponds to the linear solution. The linear solution used by Dansereau *et al*. [12] does not consider any constraint to obtain the homographies between each viewpoint and the calibration grid pose, *i.e*. for a Lytro Illum camera one computes $M \times 15 \times 15$ homographies where M corresponds to the number of calibration grid poses. On the other hand, the proposed method exploits the geometry of the viewpoint camera array to estimate a parametric homography matrix that characterizes the SPC for each calibration grid pose, *i.e*. M homographies are computed. Additionally, in Dansereau *et al*. [12], the principal point shift is assumed to be zero on the linear solution and is only estimated during the nonlinear refinement. The more accurate representation of the viewpoint camera array by the calibration proposed allows to obtain an initial solution that is closer to the final one.

Finally, let us evaluate the quality of the estimated poses and of the distortion model. For the estimated poses, one considered an image that corresponds to the mean of the intensity values after warping all VIs using the homography matrix estimated from LFIM entries for all calibration grid poses. The images for Dataset Illum-1 for the initial and final stages of the calibration process are depicted in Figure 7. Notice that in the final stage of the calibration, the edges of the calibration grid are not blurred.



(a) All imaged chessboard poses (b) Merged after linear solution (c) Merged after nonlinear solution

Fig. 7: Mean intensity values for all VIs warped using the homography matrix estimated from LFIM entries for all 10 calibration grid poses for Dataset Illum-1. **(a)** depicts the calibration pattern limits for the different calibration grid poses without homography correction. **(b)** depicts the images obtained for the initial linear solution and **(c)** depicts the images obtained for the nonlinear refinement with distortion estimation.

For the distortion model, one has rectified the lightfield of a scene that was not considered for the calibration using the distortion parameters estimated with the calibration proposed and *Dans13* [12]. The two approaches behave similarly in rectifying the straight lines in the foreground of the scene (Supp. Material E).

C. Viewpoint Camera Array

Figure 1 shows the raw image, reconstructed structure and the viewpoint camera array that characterizes a Lytro Illum camera. The reconstruction, Figures 1.b and 1.c, is obtained from disparities estimated with the structure tensor [1], which are converted to depth values, in metric units, based on the calibration parameters. The calibration parameters were extracted from the LFIM obtained at the final stage of the calibration procedure (Section V-B) using a subset of 10 poses of Illum-1.

The characterization of the viewpoint camera array is presented in Table V and depicted in Figure 1.c. Notice that the viewpoint cameras are virtual cameras so the properties associated with this camera array like baseline, scale factor and principal point shift will vary with different zoom and focus settings of the SPC.

Table V shows that the viewpoint camera array for Lytro Illum has a scale factor and baseline greater than the 1^{st} generation Lytro. The estimated unitary baseline length for the Lytro Illum is 0.52 mm and the maximum baseline length considering the 15×15 viewpoint cameras is 7.33 mm. Thus, the unitary baseline for the Lytro Illum is 1.41 times higher than the 1^{st} generation counterpart. If we consider the camera arrays, the maximum baseline for the Lytro Illum is 2.46 times higher.

The increased scale factor is justified by the higher spatial resolution of the raw (assuming the sensor size remains constant). Notice also the non-zero estimate for the principal point

$\mathbf{P}^{ij}$	k_u	k_v	u_0	v_0	Δx_0	Δy_0	Δu_0	Δv_0
Dataset B [12]	545.84	547.10	188.94	189.03	-0.00027	-0.00026	0.51	0.49
Illum-1 10 poses	841.55	840.40	310.76	214.68	-0.00036	-0.00038	0.28	0.29

TABLE V: Parameters to describe the viewpoint camera arrays of commercially available SPCs: Lytro Illum and 1^{st} generation Lytro cameras.

shift that defines a plane in focus Ω positioned approximately at 1.09 m. This estimate confirms that the principal point is different for each viewpoint camera and consequently the epipolar geometry for the SPC corresponds to the one defined in Section IV-B.

VII. Conclusions

In this work, we defined the mapping between the model of a SPC, the LFIM $\mathbf{H}$, and a viewpoint pinhole camera array described by a parametric projection matrix $\mathbf{P}^{ij}$. These mappings show that the viewpoint cameras differ on the location of their projection centers and on their principal points. Additionally, the EPI geometry described by the viewpoint camera array allows to define a zero disparity plane, the main lens world focal plane.

The viewpoint camera array model is used to define a linear solution for the SPC that considers two steps: i) a DLT calibration to obtain the parameters that describe the viewpoint homography matrix from point correspondences $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$, (ii) and a strategy to decompose the viewpoint homography matrix into intrinsic and extrinsic parameters based on a parametric representation of the image of the absolute conic. The linear solution is then refined using a nonlinear optimization. This strategy outperforms state of the art calibration procedures at the linear solution stage of the calibration process. This is the first work capable of estimating the principal point shift in the linear solution.

A similar calibration procedure can be performed using the MIs with the appropriate modifications. Thus, in terms of future work, we want to evolve this calibration procedure to work from features on microlenses in the raw image. Namely, considering alternative features like lines, as described in Bok *et al.* [13], since the MIs are unfocused for SPCs.

References

[1] S. G. Marto, N. B. Monteiro, J. P. Barreto, and J. A. Gaspar, “Structure from plenoptic imaging,” in *Joint IEEE International conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2017, pp. 338–343.

[2] M. Levoy and P. Hanrahan, “Light field rendering,” in *Proceedings of the International conference on computer graphics and interactive techniques (SIGGRAPH)*, vol. 96, no. 30. ACM, 1996, pp. 31–42.

[3] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, “The lumigraph,” in *Proceedings of the International conference on computer graphics and interactive techniques (SIGGRAPH)*, vol. 96, no. 30. ACM, 1996, pp. 43–54.

[4] D. Dansereau and 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.

[5] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 606–619, 2014.

[6] R. Ng, “Digital light field photography,” Ph.D. dissertation, Stanford University, 2006.

[7] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[8] C. Perwass and L. Wietzke, “Single lens 3d-camera with extended depth-of-field,” in *Proceedings of SPIE, Human Vision and Electronic Imaging XVII*, vol. 8291. International Society for Optics and Photonics, 2012, p. 829108.

[9] I. Ihrke, J. Restrepo, and L. Mignard-Debise, “Principles of light field imaging: Briefly revisiting 25 years of research,” *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 59–69, 2016.

[10] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, “Light field image processing: An overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.

[11] T. Georgiev and A. Lumsdaine, “Reducing plenoptic camera artifacts,” in *Computer Graphics Forum*, vol. 29, no. 6. Wiley Online Library, 2010, pp. 1955–1968.

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

[13] Y. Bok, H.-G. Jeon, and I. S. Kweon, “Geometric calibration of micro-lens-based light field cameras using line features,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 287–300, 2017.

[14] N. Zeller, F. Quint, and U. Stilla, “From the calibration of a light-field camera to direct plenoptic odometry,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1004–1019, 2017.

[15] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, “Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3280–3287.

[16] Y. Luan, X. He, B. Xu, P. Yang, and G. Tang, “Automatic calibration method for plenoptic camera,” *Optical Engineering*, vol. 55, no. 4, p. 043111, 2016.

[17] N. B. Monteiro, S. Marto, J. P. Barreto, and J. Gaspar, “Depth range accuracy for plenoptic cameras,” *Computer Vision and Image Understanding*, vol. 168, pp. 104–117, 2018.

[18] C. Hahne, A. Aggoun, S. Haxha, V. Velisavljevic, and J. C. Fernández, “Baseline of virtual cameras acquired by a standard plenoptic camera setup,” in *2014 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2014, pp. 1–3.

[19] C. Birklbauer and O. Bimber, “Panorama light-field imaging,” *Computer Graphics Forum*, vol. 33, no. 2, pp. 43–52, 2014.

[20] M. D. Grossberg and S. K. Nayar, “The raxel imaging model and ray-based calibration,” *International Journal of Computer Vision*, vol. 61, no. 2, pp. 119–137, 2005.

[21] R. C. Bolles, H. H. Baker, and D. H. Marimont, “Epipolar-plane image analysis: An approach to determining structure from motion,” *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[22] P. David, M. Le Pendu, and C. Guillemot, “White lenslet image guided demosaicing for plenoptic cameras,” in *19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[23] Y. Abdel-Aziz, “Karara. hm 1971. direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry,” in *Proceedings ASP/VI Symp. On Close-Range Photogrammetry*, 1971, pp. 1–17.

[24] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[25] H. Lutkepohl, “Handbook of matrices.” *Computational Statistics and Data Analysis*, vol. 2, no. 25, p. 243, 1997.

[26] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, 2000.

[27] Q.-T. Luong and O. D. Faugeras, “Self-calibration of a moving camera from point correspondences and fundamental matrices,” *International Journal of computer vision*, vol. 22, no. 3, pp. 261–289, 1997.

[28] O. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993.

[29] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*. Siam, 2000, vol. 1.

Nuno Barroso Monteiro received the B.E. degree and Masters degree in Biomedical Engineering from Instituto Superior Tcnico (IST), Portugal, in 2007 and 2009, respectively. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering from IST. His research interests include multiview stereo reconstruction, lightfield camera model and calibration.

Joao P. Barreto (M'99) received the Licenciatura and Ph.D. degrees from the University of Coimbra, in 1997 and 2004, respectively. From 2003 to 2004, he was a Post-doctoral Researcher with the University of Pennsylvania, Philadelphia. He has been a Professor with the University of Coimbra, since 2004, where he is also a Senior Researcher with the Institute for Systems and Robotics. His current research interests include different topics in 3D computer vision, with a special emphasis in robotics and medicine. He is the author of more than 80 peer-reviewed publications and recipient of several distinctions and awards including a Google Faculty Research award and 5 Outstanding Reviewer Awards. He has served as Area Chair in the most prestigious conferences in computer vision (ICCV, ECCV, CVPR) and he is currently Associate Editor for Computer Vision and Image Understanding, Image and Vision Computing and Journal of Mathematical Imaging and Vision.

José António Gaspar received his Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Tcnico (IST), Portugal, in 2003. He is currently an Auxiliary Professor at IST, University of Lisbon, and a Researcher at the Computer Vision Laboratory (VisLab), Institute for Systems and Robotics (ISR). His research interests are in Computer and Robot Vision, Robotics and Control. The most recent research has been focused on modelling Non-Conventional Cameras.

Standard Plenoptic Cameras Mapping to Camera Arrays and Calibration based on DLT - Supplementary Material

Nuno Barroso Monteiro, Joao P. Barreto, José António Gaspar

Introduction to the Supplementary Material

The supplementary material deduces some of the formulas used in the main paper and provides more insights regarding the mapping between a standard plenoptic camera (SPC) and the viewpoint camera array. Namely, one explains the parameterization of the rays by a point and a direction and the influence of re-parameterization on the lightfield intrinsics matrix (LFIM) $\mathbf{H}$ in Supp. Material A. In Supp. Material B, one presents the location of the viewpoint projection centers and the restriction to consider the viewpoint cameras as pinholes. The reduction of the number of non-zero entries in the LFIM by considering the viewpoint projection centers location is explained in Supp. Material C. These sections are the basis for the LFIM representation (2) with 8 non-zero entries and for the simpler notation of the viewpoint camera mapping from the LFIM (6). The supplementary material also addresses some practical aspects regarding the homography estimation (Supp. Material D) and presents additional results regarding the quality of the distortion model (Supp. Material E) and the precision associated with each entry of the LFIM with the number of poses (Supp. Material F). For a complete understanding of the notation, please refer to the main article.

A. Ray Parameterization and Re-Parameterization

Consider a lightfield in the object space $L_\Pi(q,r,u,v)$ acquired by a plenoptic camera with the plane Ω in focus (Figure A.1). $L_\Pi(q,r,u,v)$ is a set of rays, where each ray $\tilde{\mathbf{\Psi}}_\Pi = [q,r,u,v,1]^T$ is parameterized using a point (q,r) on a plane Π and a direction (u,v) defined in metric units [1]. This lightfield is mapped to the lightfield in the image space $L(i,j,k,l)$ by the LFIM $\mathbf{H}_\Pi$ introduced by Dansereau *et al.* [2]:

$$\tilde{\mathbf{\Psi}}_\Pi = \mathbf{H}_\Pi \; \tilde{\mathbf{\Phi}} \quad , \tag{A.1}$$

*This work was supported by the Portuguese Foundation for Science and Technology (FCT) projects (grant numbers UID/EEA/50009/2019, PD/BD/105778/2014, and PINFRA/22084/2016) and the European Commission project ORIENT (ERC/2016/693400). J. P. Barreto thanks FCT and COMPETE2020 for generous funding under grant PTDC/EEIAUT/3024/2014.

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

*J. P. Barreto is with the Institute for Systems and Robotics, University of Coimbra, Portugal, jpbar@isr.uc.pt

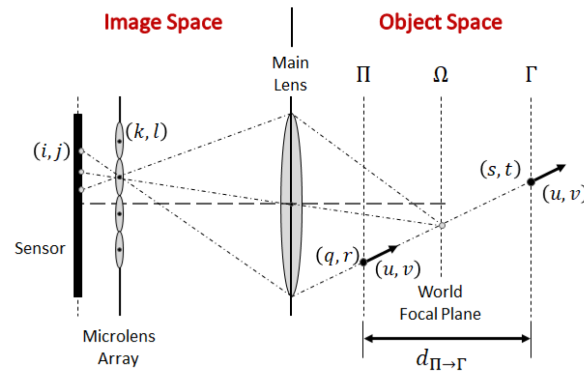


Fig. A.1: Geometry of a SPC. The lightfield in the image space is parameterized using pixels and microlenses indices while the lightfield in the object space is parameterized using a point and a direction. The lightfield in the object space can be parameterized on an arbitrary plane regardless of the original plane Ω in focus.

where $\tilde{\mathbf{\Phi}} = [i,j,k,l,1]^T$ corresponds to a ray that is parameterized by pixels (i,j) and microlenses (k,l) indices and

$$\mathbf{H}_\Pi = \begin{bmatrix} h_{qi} & 0 & h_{qk} & 0 & h_q \\ 0 & h_{rj} & 0 & h_{rl} & h_r \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad . \tag{A.2}$$

This mapping allows writing the positions (q,r) and the directions (u,v) as affine mappings on the pixels (i,j) and microlenses (k,l) indices.

On the other hand, the lightfield in the object space $L_\Pi(q,r,u,v)$ can be redefined on another plane Γ by shifting the parameterization plane Π along the optical axis of the SPC, *i.e.* along the normal to the plane Π. Assuming that Γ is at a distance $d_{\Pi\to\Gamma}$ from Π, the re-parameterization [3] is defined as

$$\tilde{\mathbf{\Psi}}_\Gamma = \mathbf{D} \; \tilde{\mathbf{\Psi}}_\Pi \tag{A.3}$$

where

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & d_{\Pi\to\Gamma} & 0 & 0 \\ 0 & 1 & 0 & d_{\Pi\to\Gamma} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad . \tag{A.4}$$

Note that $\mathbf{D}$ maps a ray $\tilde{\mathbf{\Psi}}_\Pi$ to a ray $\tilde{\mathbf{\Psi}}_\Gamma = [s,t,u,v,1]^T$ representing a ray passing on a point (s,t) on plane Γ with a direction (u, v). Notice that $\mathbf{D}$ changes the camera coordinate system origin but does not change the directions (u, v).

Mapping the lightfield in the object space $L_\Pi(q, r, u, v)$ to the lightfield in the image space $L(i, j, k, l)$ by (A.1), one has

$$\tilde{\mathbf{\Psi}}_\Gamma = \mathbf{D}\ \mathbf{H}_\Pi\ \tilde{\mathbf{\Phi}} \quad . \tag{A.5}$$

The intrinsic matrix $\mathbf{H}_\Gamma = \mathbf{D}\ \mathbf{H}_\Pi$ maps the lightfield in the image space $L(i, j, k, l)$ and the lightfield in the object space $L_\Gamma(s, t, u, v)$.

B. Viewpoint Pinhole Constraint

In this section, we show that the SPC model [2], represented in the form (A.2), is equivalent to an array of parallel viewpoint cameras.

From the previous section, let us consider the lightfield in the object space whose rays are parameterized at a plane Π using a point $[q, r, 0]^T$ and a direction $[u, v, 1]^T$ (Figure A.1). The LFIM $\mathbf{H}_\Pi$ (A.2) maps the rays in the image space $\mathbf{\Phi}$ to the rays in the object space $\mathbf{\Psi}_\Pi$ [2] by (A.1). For a viewpoint or sub-aperture camera, the pixel coordinates (i, j) are fixed and are considered as parameters. Hence, for a viewpoint camera, the positions (q, r) and the directions (u, v) are affine mappings only on the microlens coordinates (k, l), namely

$$\begin{cases} q(k;\ i, \mathbf{H}_\Pi) = h_{qk}\, k + h_{qi}\, i + h_q \\ r(l;\ j, \mathbf{H}_\Pi) = h_{rl}\, l + h_{rj}\, j + h_r \\ u(k;\ i, \mathbf{H}_\Pi) = h_{uk}\, k + h_{ui}\, i + h_u \\ v(l;\ j, \mathbf{H}_\Pi) = h_{vl}\, l + h_{vj}\, j + h_v \end{cases} \tag{A.6}$$

where the LFIM $\mathbf{H}_\Pi$ is also considered as a parameter. To simplify the notation, we will not include the parameters $(i, j, \mathbf{H}_\Pi)$ in the following expressions.

A ray captured by a SPC and parameterized by (i, j, k, l) intersects the plane Π at point $\mathbf{p}(k, l) = [q(k),\ r(l),\ 0]^T$ with a direction $\mathbf{n}(k, l) = [u(k),\ v(l),\ 1]^T$. This allows to define an arbitrary point $\mathbf{c}(k, l, \lambda) = [x, y, z]^T$ along the ray [4] as

$$\mathbf{c}(k, l, \lambda) = \mathbf{p}(k, l) + \lambda\ \mathbf{n}(k, l)\ \ ,\ \lambda \in \mathbb{R} \quad . \tag{A.7}$$

Note that by sweeping the range of (k, l) in (A.7) with $\lambda = 0$, one samples an area of the plane Π through which pass all the viewpoint imaging rays. In addition, by sweeping (i, j), one obtains all the viewpoints, and therefore all rays that can be imaged by the SPC. Finally, sweeping λ, allows representing all world points within the field of view of the SPC.

The location of the projection centers of an optical setup is defined by its caustic surface, which is the loci of singularities in the flux density [4], [5]. The convergence of the rays captured by a camera at a single point, *i.e.* a unique projection center, is considered a degenerate configuration of the caustic surface (point caustic) [4]. Although there are many techniques to derive the caustic surface, in this work, we will consider the Jacobian method [5].

The caustic surface is defined at the points in the object space where the ray to image mapping (A.7) is singular, *i.e.* the mapping from (k, l, λ) to (x, y, z) is singular. The singularities occur at the set of points where the Jacobian matrix of the transformation does not have full rank, *i.e.* points that make the determinant of the Jacobian vanish $\det\left(\mathbf{J}\left(\mathbf{c}(k, l, \lambda)\right)\right) = 0$. Solving the vanishing constraint one obtains two solutions for λ:

$$\lambda_1 = -\frac{h_{qk}}{h_{uk}} \quad \vee \quad \lambda_2 = -\frac{h_{rl}}{h_{vl}} \quad . \tag{A.8}$$

Replacing λ_1 or λ_2 in equation (A.7) identifies the caustic profile for the viewpoint camera. The caustic profile of a single viewpoint consists of a line with (i) unique (x, z) and variable y components if $\lambda = \lambda_1$ or (ii) unique (y, z) and variable x components if $\lambda = \lambda_2$. In case $\lambda_1 \neq \lambda_2$ the viewpoint is a non-central camera. The viewpoint camera corresponds to a central camera, *i.e.* a camera with a unique projection center, if and only if $\lambda_1 = \lambda_2$ which imply the model parameters relation

$$\frac{h_{qk}}{h_{uk}} = \frac{h_{rl}}{h_{vl}} \quad . \tag{A.9}$$

Assuming this constraint and replacing λ in (A.7), expanded by the expressions in (A.6), the location of the viewpoint projection center for a viewpoint camera (i, j) is given by

$$\mathbf{p}_c = \begin{bmatrix} h_q - \frac{h_{qk}}{h_{uk}} h_u + i\left(h_{qi} - \frac{h_{qk}}{h_{uk}} h_{ui}\right) \\ h_r - \frac{h_{rl}}{h_{vl}} h_v + j\left(h_{rj} - \frac{h_{rl}}{h_{vl}} h_{vj}\right) \\ -\frac{h_{qk}}{h_{uk}} \end{bmatrix} \quad . \tag{A.10}$$

Furthermore, considering all viewpoint cameras that can be defined for a SPC, the SPC originates a co-planar grid of equally spaced projection centers. Notice that the pixels (i, j) only affect the x- and y-components of the projection centers while the z-component of the projections centers is always the same.

C. Reducing the Parameters of the LFIM Parameterization

The LFIM has 12 non-zero entries (A.2) that help introducing and explaining the model but some parameters can be avoided by considering them on the extrinsic parameters.

Considering the parameterization plane Π (Figure A.1) for the origin of the different rays $\tilde{\mathbf{\Psi}}_\Pi = [q, r, u, v, 1]^T$ in the object space, an arbitrary point is defined as $[x, y, z]^T = [q, r, 0]^T + \lambda\ [u, v, 1]^T$, $\lambda \in \mathbb{R}$ [4]. The re-parameterization of the rays in the object space to the plane Γ (A.3) corresponds to a shift along the z-axis of the camera coordinate system, which results in $[x, y, z_\Gamma]^T = [s, t, 0]^T + \lambda\ [u, v, 1]^T$ where $s = q + u\, d_{\Pi\to\Gamma}$, $t = r + v\, d_{\Pi\to\Gamma}$, and $z_\Gamma = z - d_{\Pi\to\Gamma}$. Thus, the re-parameterization is redundant with the z-translation of the extrinsic parameters. Assuming that the plane Γ corresponds to the plane containing the viewpoint projection centers at $d_{\Pi\to\Gamma} = -h_{qk}/h_{uk}$ (Supp. Material B), one obtains a LFIM $\mathbf{H}_\Gamma$ with 10 non-zero entries

$$\mathbf{H}_\Gamma = \begin{bmatrix} h_{si} & 0 & 0 & 0 & h_s \\ 0 & h_{tj} & 0 & 0 & h_t \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad . \tag{A.11}$$

Furthermore, extending the definition of the point (s,t) to consider the lightfield coordinates in the image space and redefining x and y as $x_\Gamma = x - h_s$ and $y_\Gamma = y - h_t$, one obtains $[x_\Gamma, y_\Gamma, z_\Gamma]^T = [h_{si}\ i, h_{tj}\ j, 0]^T + \lambda\ [u, v, 1]^T$. Hence, as identified by Dansereau *et al*. [2], the entries h_s and h_t are redundant with the (x,y)-translational components of the extrinsic parameters. Thus, removing the redundant entries, one obtains a LFIM $\mathbf{H}_\Gamma$ with 8 non-zero entries

$$\mathbf{H}_\Gamma = \begin{bmatrix} h_{si} & 0 & 0 & 0 & 0 \\ 0 & h_{tj} & 0 & 0 & 0 \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad . \tag{A.12}$$

This is the LFIM representation that is considered on the main paper. Considering this representation for the LFIM, the viewpoint projection centers location (A.10) reduces to

$$\mathbf{p}_c = \begin{bmatrix} i\ h_{si} \\ j\ h_{tj} \\ 0 \end{bmatrix} \quad , \tag{A.13}$$

which is the one considered on the main paper, namely $\mathbf{t}^{ij} = -\mathbf{p}_c$.

D. Practical Aspects of the Homography Estimation

The parametric homography matrix (10) can be estimated using a direct linear transformation (DLT) [6]. Denoting the unknown parameters of the homography matrix as $\mathbf{b}$ and the observation matrix for each point correspondence $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$ as $\mathbf{M}_n$ in equation (12), one has

$$\underbrace{\left(\tilde{\mathbf{m}}^T \otimes [\tilde{\mathbf{q}}]_\times\right) \mathbf{T}}_{\mathbf{M}_n} \underbrace{\begin{bmatrix} \mathbf{h}^0 \\ \mathbf{a}^{ij} \end{bmatrix}}_{\mathbf{b}} = \mathbf{0}_{3\times 1} \quad . \tag{A.14}$$

Considering an observation matrix $\mathbf{M}$ obtained from stacking the matrices $\mathbf{M}_n$ of each pair $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$, the solution corresponds to a non-zero vector in the null space of $\mathbf{M}$. Since the projection equation is defined up to a scale factor, one should constraint the solution to $\|\mathbf{b}\|^2 = 1$ leading to the following optimization problem

$$\begin{aligned} \hat{\mathbf{b}} &= \arg\min_{\mathbf{b}}\ \|\mathbf{M}\,\mathbf{b}\|^2 \\ \text{s.t.} &\quad \|\mathbf{b}\|^2 = 1 \end{aligned} \quad . \tag{A.15}$$

In order to obtain an estimate for the homography matrix (11), one should have present two practical aspects:

a) Data Normalization: For a DLT it is crucial to normalize the data in order to improve the condition number of the matrix $\mathbf{M}^T\mathbf{M}$ [7]. Thus, one should consider a translation of the image points and the points in the object space so that their centroids are at the origin and the average distances to the origin are equal to $\sqrt{2}$ and $\sqrt{3}$ [8], respectively.

b) Computing a Solution in case of a Large Number of Features: In order to build an over-determined system, having a least squares solution, one should use each projection $\mathbf{q}$ observed in each viewpoint camera for a given point $\mathbf{m}$. Therefore, assuming a plenoptic camera with N pixels within each microlens, a point in the object space generates N pairs $(\tilde{\mathbf{m}}, \tilde{\mathbf{q}})$, and consequently $2N$ equations. Normally, in a calibration procedure, ones uses a calibration grid, with K feature points, that is observed in C different poses. This leads to a "tall" observation matrix $\mathbf{M}$ with $L = 2N \times K \times C$ rows and 20 columns, *i.e*. one has a high number of observations compared with the number of parameters to estimate. Consequently, using a singular value decomposition (SVD) to obtain the solution to the optimization problem (A.15) is troublesome since this decomposition needs to compute the square matrix $\mathbf{M}^T\mathbf{M}$ with size $L \times L$ which requires a prohibitive storage space. Hence, a solution is to perform a QR-Decomposition [9] of the observation matrix $\mathbf{M} = \mathbf{Q}\begin{bmatrix}\mathbf{V} & \mathbf{0}_{(L-20)\times 20}\end{bmatrix}^T$ where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{V}$ is an upper triangular matrix with size 20×20. This allows to rewrite the optimization problem (A.15) as

$$\begin{aligned} \hat{\mathbf{b}} &= \arg\min_{\mathbf{b}}\ \|\mathbf{V}\,\mathbf{b}\|^2 \\ \text{s.t.} &\quad \|\mathbf{b}\|^2 = 1 \end{aligned} \quad , \tag{A.16}$$

which can be solved using SVD.

E. Distortion Model

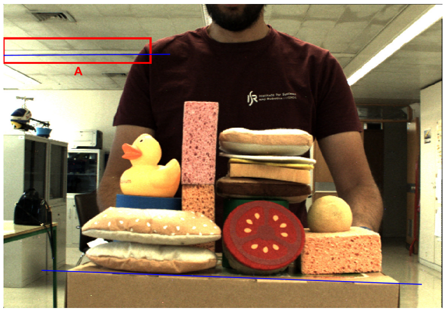
The distortion model is evaluated rectifying the lightfield of a scene that was not considered for the calibration using the distortion parameters estimated with the calibration proposed in Section V and the calibration procedure of Dansereau *et al*. [2] (denoted as *Dans13*). The central VI of the rectified lightfield considering the results of the calibration on a subset of 10 poses for Dataset Illum-2 is presented in Figure A.2. The radial distortion considered allows to rectify correctly the straight lines in the foreground of the scene (Figure A.2.b-c). Notice that for *Dans13* [2] (Figure A.2.d), the straight lines in the background are distorted in the rectification. Nonetheless, the rectification using the parameters estimated with the calibration proposed allows to maintain straight lines in the background and in the foreground (Figure A.2.e).

F. Additional Results on LFIM Parameters Estimation

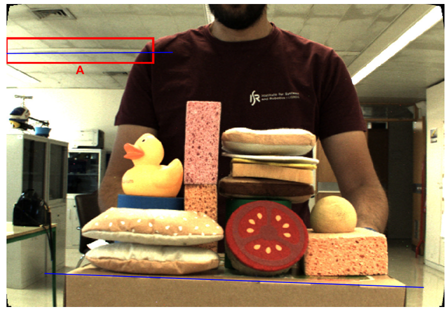
In order to evaluate the precision of the calibration, we repeated 20 times the calibration procedure. Each calibration involves $k = 2, \ldots, 20$ pattern poses, randomly selected from the full calibration dataset. The calibration procedure proposed in Section V is compared with the methodology [2] (denoted as *Dans13*).

The full calibration dataset is acquired with a Lytro Illum camera using two calibration grids with different sizes: 8×6 grid of 211×159 mm with approximately 26.5 mm cells (denoted as Illum-1), and 20×20 grid of 121.5×122 mm with approximately 6.1 mm cells (denoted as Illum-2). Each dataset acquired is composed of 66 fully observable poses of the calibration pattern. Care was taken to avoid changing the focal settings of the camera.

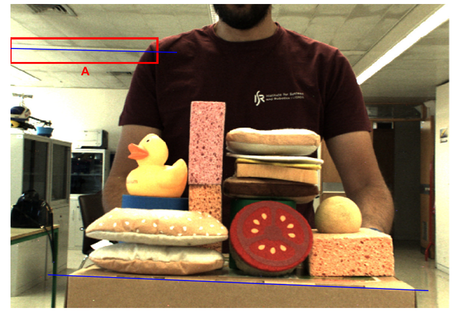
The mean and standard deviation obtained for each parameter of LFIM for Datasets Illum-1 and Illum-2 are depicted in



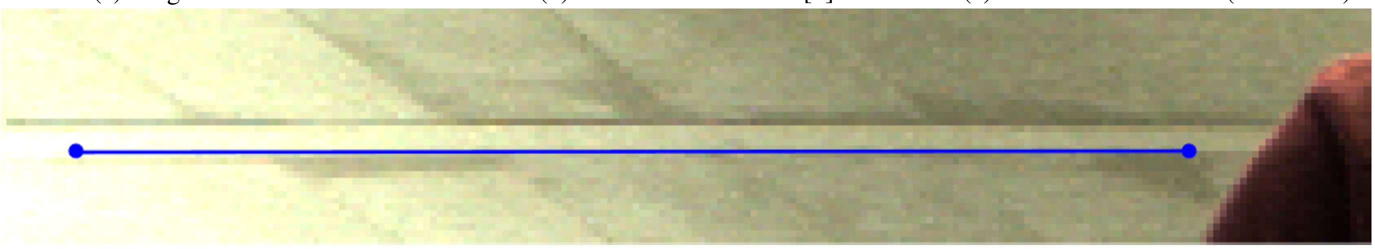
(a) Original central VI



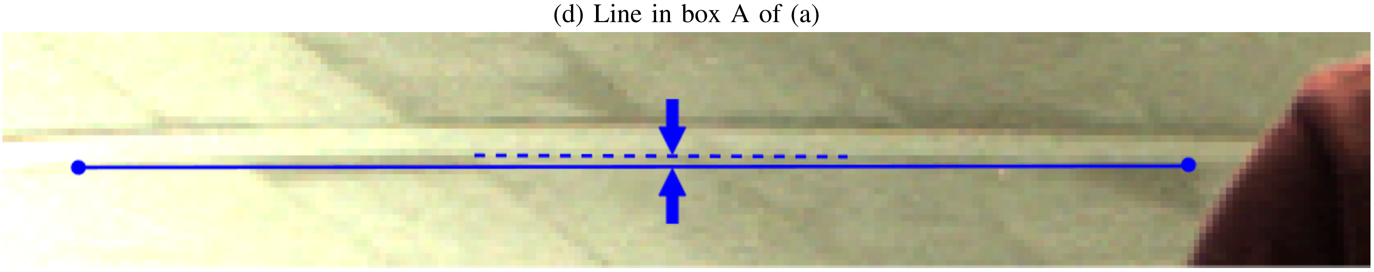
(b) Rectified central VI [2]



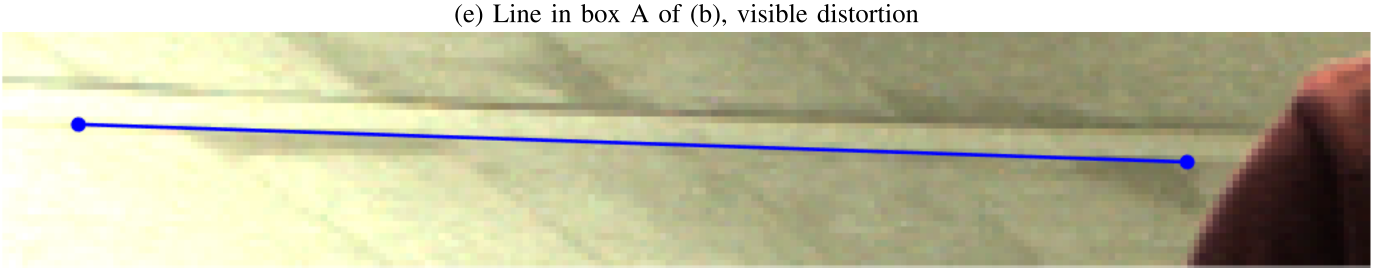
(c) Rectified central VI (Section V)



(d) Line in box A of (a)



(e) Line in box A of (b), visible distortion



(f) Line in box A of (c), lesser visible distortion

Fig. A.2: Distortion rectification using the distortion parameters estimated with *Dans13* [2] (**(b)** and **(e)**) and the calibration proposed (**(c)** and **(f)**) for the Dataset Illum-2. **(a)** depicts the original central VI while **(d)-(f)** correspond to zooms of the red boxes A. Blue rulings were added to aid in the visual confirmation of the straight lines after rectification.

Figures A.3 and A.4. Notice that the calibration *Dans13* [2] obtains a LFIM with 12 non-zero entries while the method proposed has 8 non-zero entries. For comparing the parameters, we transformed the LFIM obtained by *Dans13* as defined in Supp. Material C. Figures A.3 and A.4 show that a minimum of 8-9 poses are needed for a precise estimation of the LFIM parameters. Namely, to have a deviation smaller than 3% of the mean value, one needs 9 poses using *Dans13* [2] and 8 poses using the proposed calibration for Dataset Illum-1. For Dataset Illum-2, one needs 9 poses using *Dans13* [2] and the proposed calibration.

Let us also consider the statistical analysis of the difference between the estimates at the initial and final stages of the calibration process for each of the entries of the LFIM. The mean and standard deviation values for Dataset Illum-1 and Illum-2 are depicted in Figures A.5 and A.6. These figures show that the calibration proposed allows to obtain an initial solution that is closer to the solution at the final stage of the calibration procedure. Namely, the proposed calibration allows to estimate more precisely the entries related with the baseline and the principal point shift (Figure A.5.a-b and A.6.a-b). For the remaining entries, the performance is similar for both calibration methods.

References

[1] N. B. Monteiro, S. Marto, J. P. Barreto, and J. Gaspar, "Depth range accuracy for plenoptic cameras," *Computer Vision and Image Understanding*, vol. 168, pp. 104–117, 2018.

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

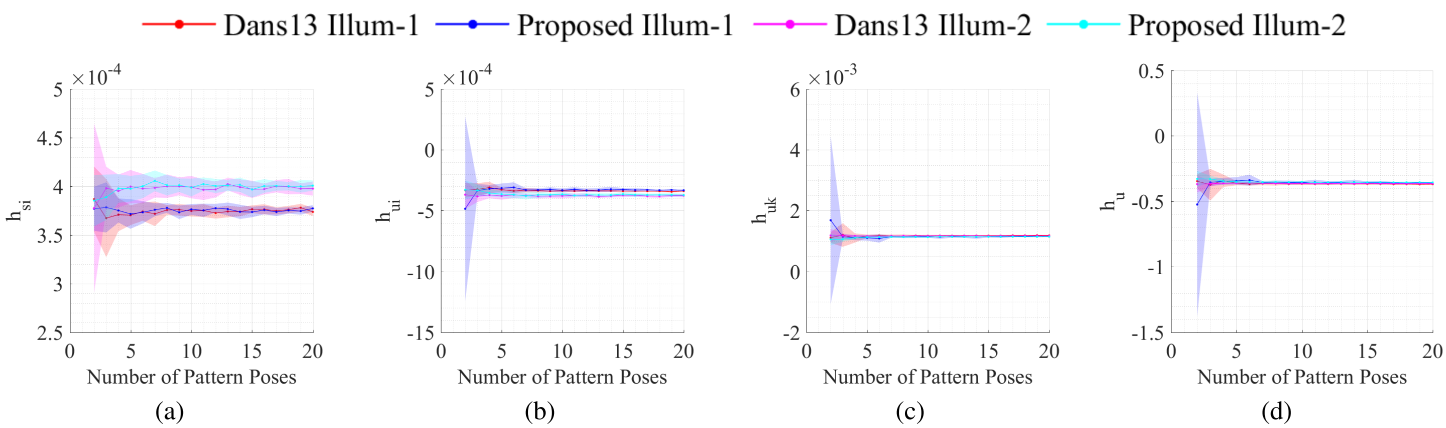


Fig. A.3: Precision of the LFIM parameters after nonlinear refinement with distortion estimation using the calibration proposed (in blue and cyan for dataset Illum-1 and Illum-2, respectively) and the calibration *Dans13* [2] (in red and magenta for dataset Illum-1 and Illum-2, respectively): h_{si} **(a)**, h_{ui} **(b)**, h_{uk} **(c)**, and h_u **(d)**. The mean values are represented by the solid lines and the standard deviation by the shaded areas.

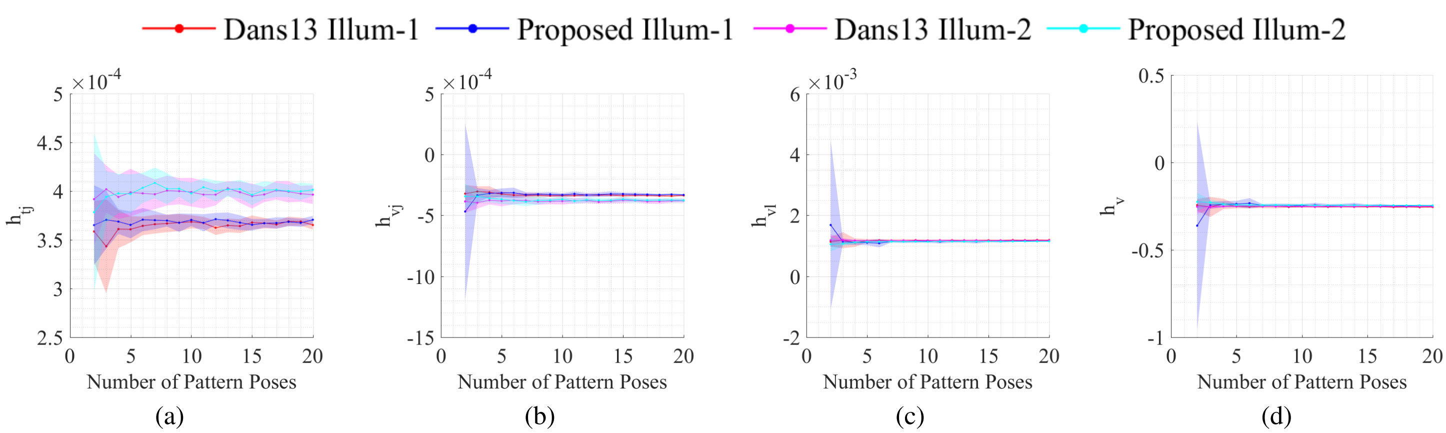


Fig. A.4: Precision of the LFIM parameters after nonlinear refinement with distortion estimation using the calibration proposed (in blue and cyan for dataset Illum-1 and Illum-2, respectively) and the calibration *Dans13* [2] (in red and magenta for dataset Illum-1 and Illum-2, respectively): h_{tj} **(a)**, h_{vj} **(b)**, h_{vl} **(c)**, and h_v **(d)**. The mean values are represented by the solid lines and the standard deviation by the shaded areas.

[3] C. Birklbauer and O. Bimber, "Panorama light-field imaging," *Computer Graphics Forum*, vol. 33, no. 2, pp. 43–52, 2014.

[4] M. D. Grossberg and S. K. Nayar, "The raxel imaging model and ray-based calibration," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 119–137, 2005.

[5] D. G. Burkhard and D. L. Shealy, "Flux density for ray propagation in geometrical optics," *JOSA*, vol. 63, no. 3, pp. 299–304, 1973.

[6] Y. Abdel-Aziz, "Karara. hm 1971. direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry," in *Proceedings ASP/VI Symp. On Close-Range Photogrammetry*, 1971, pp. 1–17.

[7] R. I. Hartley, "In defence of the 8-point algorithm," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*. IEEE, 1995, pp. 1064–1070.

[8] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[9] J. E. Gentle, *Numerical linear algebra for applications in statistics*. Springer Science & Business Media, 2012.

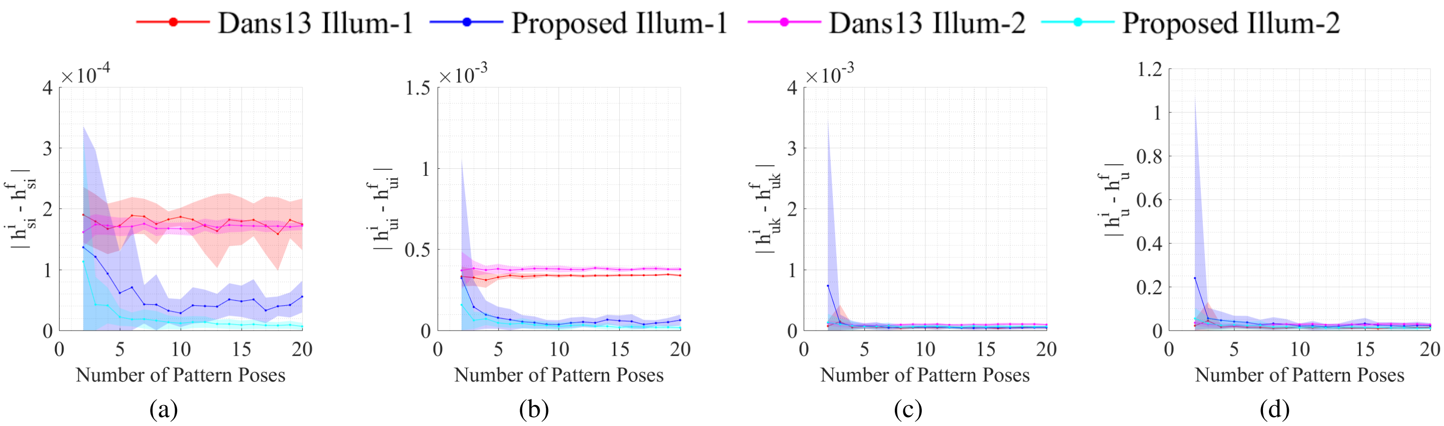


Fig. A.5: Difference between the estimated LFIM parameters at the initial and final stages of the calibration proposed (in blue and cyan for dataset Illum-1 and Illum-2, respectively) and the calibration *Dans13* [2] (in red and magenta for dataset Illum-1 and Illum-2, respectively): h_{si} **(a)**, h_{ui} **(b)**, h_{uk} **(c)**, and h_u **(d)**. The mean values are represented by the solid lines and the standard deviation by the shaded areas.

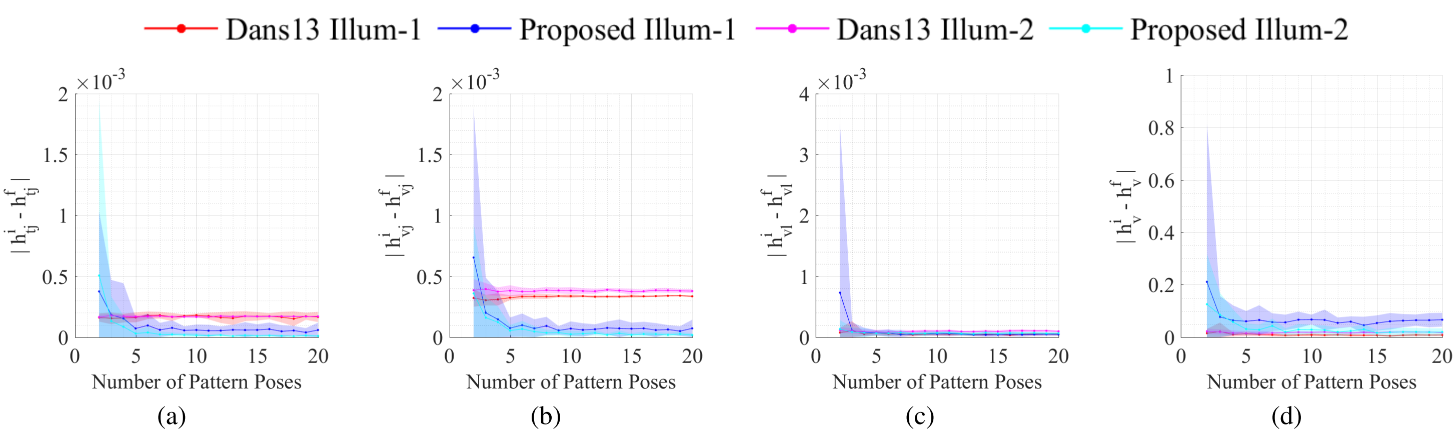


Fig. A.6: Difference between the estimated LFIM parameters at the initial and final stages of the calibration proposed (in blue and cyan for dataset Illum-1 and Illum-2, respectively) and the calibration *Dans13* [2] (in red and magenta for dataset Illum-1 and Illum-2, respectively): h_{tj} **(a)**, h_{vj} **(b)**, h_{vl} **(c)**, and h_v **(d)**. The mean values are represented by the solid lines and the standard deviation by the shaded areas.