Locally Affine Light Fields as Direct Measurements of Depth

Simão Pedro da Graça Oliveira Marto
smarto@isr.tecnico.ulisboa.pt
Nuno Barroso Monteiro
nmonteiro@isr.tecnico.ulisboa.pt
José António Gaspar
jag@isr.tecnico.ulisboa.pt

Institute for Systems and Robotics
Instituto Superior Técnico
University of Lisbon, Portugal

Abstract

Light field imaging allows discriminating object radiance according to multiple viewing directions. We introduce the minimal light field representation from which depth can be extracted, the affine light field, which is a first order approximation. One setup to acquire one globally affine light field is proposed. Consequently, we show how Dansereau Bruton's gradient based reconstruction method [1] can be derived from the locally affine light field assumption.

1 Introduction

Light field cameras, sometimes called plenoptic cameras, have been introduced recently to the consumer market [4]. They 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.

A light field image is usually represented by the 4D plenoptic function [4], but it can be seen as a collection of 2D viewpoint images, each with a projection center slightly offset (details in section 2). This means that from a light field image it is possible to extract depth information. The only requirements is that the gradients in a viewpoint image are not null.

In this paper we introduce a minimal order approximation for a light field image which still contains depth information. It is a first order approximation due to the constraint that the gradients cannot be null. We refer to such light fields as globally or locally affine. An example setup to capture a globally affine light field is illustrated in Fig. 1. We use this approximation to derive the formula to extract depth from a light field image.

2 Light Field Camera Model

A light field image is a mapping of rays into light intensities. We make the distinction between the light field in the object space, indexed by rays in the object space, and the light field in the image space, indexed by rays in the image space. When a light field image is captured, it's in the image space, but in order to obtain metric information about a scene, the light field must first be converted into the object space.

The model proposed by Dansereau *et al.* considers a mapping between rays in the image space, sometimes referred to as raxels [4], and rays in the object space. This is the light field equivalent of an intrinsic camera model.

The rays in the object space are modelled with the two plane parameterization, see Fig. 2. Each ray is defined by its intersection with a plane (s,t) and its direction is defined by slopes (u,v) relative to the z axis. To help illustrate this parameterization, and to facilitate the understanding of the calculations in the next sections, one writes $(x,y) = (s,t) + z \cdot (u,v)$ to show how the (x,y) coordinates of a point along a ray (s,t,u,v) can be calculated from its z coordinate.

The typical construction of a light field camera is based on an array of microlenses placed between the camera main lens and the imaging sensor (usually a CMOS). The raw image extracted from the CMOS results in the so-called image in the image space after a decoding process.

In the image space, coordinates (k,l) indicate the microlens the ray passed through before sampling, and (i,j) indicate the pixel within the microlens image. Alternatively, (i,j) can be seen as selecting a viewpoint, and (k,l) as selecting a pixel within that viewpoint image. Changing (i,j) changes the projection center of the viewpoint image slightly within a plane parallel to the image plane. Another useful construct is the Epipolar Plane Image (EPI), obtained by fixing (j,l) (horizontal EPI) or fixing (i,k) (vertical EPI).

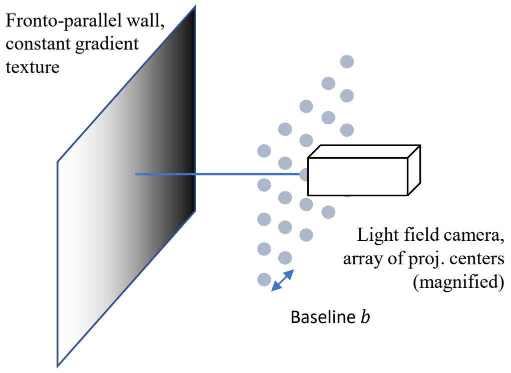

Figure 1: Setup to acquire a globally affine light field with a light field camera. The central circle represents the projection center of the central viewpoint of the light field camera. The array of circles represents the array of projection centers (not in scale) representing the other viewpoints of the light field camera.

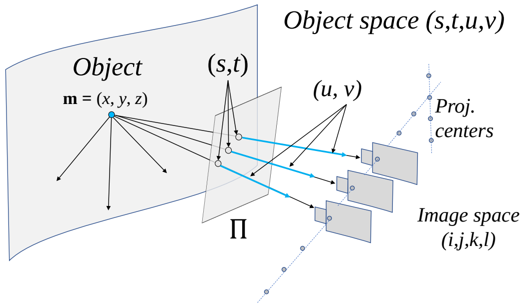

Figure 2: Two plane parameterization for rays starting from a point $\mathbf{m}$ using a point and a direction. The point (s,t) is given by the intersection with the plane Π, and the direction (u,v) with the derivative of the ray's coordinates with respect to z. The latter coordinates can also be seen as the intersection with a plane perpendicular to the first at a distance of one unit, hence the name "two plane parametrization".

The model defined in [2] takes the form of:

$$\underbrace{\begin{bmatrix} s \\ t \\ u \\ v \\ 1 \end{bmatrix}}_{\Psi} = \underbrace{\begin{bmatrix} h_{si} & 0 & h_{sk} & 0 & h_s \\ 0 & h_{tj} & 0 & h_{tl} & h_t \\ h_{ui} & 0 & h_{uk} & 0 & h_u \\ 0 & h_{vj} & 0 & h_{vl} & h_v \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} i \\ j \\ k \\ l \\ 1 \end{bmatrix}}_{\Phi} . \qquad (1)$$

The model in Eq. 1 has 8 parameters. The values in the last column of the matrix are not independent parameters, they are set by the requirement that Ψ_{center} should map to Φ_{center}. With a simple and reasonable set of assumptions, these can be reduced to just two parameters, and a meaning can be assigned to them by making an analogy to a camera array.

The first simplifying approximation is to consider the parameters referring to the horizontal and to the vertical coordinates to be equal. This is supported by the fact that, although the microlens array structure is hexagonal, a decoding algorithm can re-sample the microlenses in a square lattice, as is done by Dansereau *et al.* in [2].

Afterwards, we can move the (s,t) plane along z, such that it now includes the centres of projection of the viewpoints. The result is a new $\mathbf{H}_a$ matrix that describes the exact same camera, but has $h_{sk} = h_{tl} = 0$.

The translation in z would be compensated by an opposite translation in the extrinsic parameters.

Furthermore, the terms h_{ui} and h_{vj} describe a shift of the principal point of each viewpoint image proportional to (i,j). Since this shift can be easily removed from an image, we will consider it to be zero.

Combining all of these simplifications, we get an intrinsics matrix $\mathbf{H}_a$ with 2 parameters (apart from the 4 in the last column, which continue to be set by the requirement that Φ_{center} maps to Ψ_{center})

$$H = \begin{bmatrix} b & 0 & 0 & 0 & s_0 \\ 0 & b & 0 & 0 & t_0 \\ 0 & 0 & f^{-}1 & 0 & -c_x/f \\ 0 & 0 & 0 & f^{-}1 & -c_x/f \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad , \tag{2}$$

where the parameters $h_{si} = h_{tj}$ and $h_{uk} = h_{vl}$ are replaced by b and f^{-1}. The reason for this substitution is that these terms now have meaning in terms of a camera array: b is the baseline, or the distance between adjacent cameras; f is the focal length of the cameras. A more detailed explanation of the intrinsics matrix applied to a camera array can be found in [3].

3 Affine Light Field and Depth Estimation

The light field of a fronto-parallel plane *colored* with a gradient, Fig. 1, is the simplest scenario producing an affine light field. To show this, consider a fronto-parallel plane Π where $\mathbf{n} = (0,0,1)$, and $r = z$, such that $\mathbf{p} \in \Pi \implies \mathbf{p} \cdot \mathbf{n} = r$. The color of the plane at a point $\mathbf{p}$ in the plane Π, is given by $c(\mathbf{p}) = \mathbf{p} \cdot \mathbf{g} + c_0$, where $\mathbf{g}$ is the color gradient, and is a vector aligned with the plane.

To find out the color sampled by a ray Ψ, we find out where it hits the plane Π using the back projection equation $\left([s\ t\ 0]^T + \lambda [u\ v\ 1]^T\right) \cdot \mathbf{n} = r$. Note that λ, the parameter representing how much the ray extends before hitting Π, is actually the depth z of this plane at that point. From here on, we will use $z = \lambda$. Hence one has $z = \dfrac{r - (s,t,0) \cdot \mathbf{n}}{(u,v,1) \cdot \mathbf{n}} = r$. Combining with the camera parameterization in Eq. 1, we get the affine light field:

$$L(i,j,k,l) = l_0 + \left[a_i\ a_j\ a_k\ a_l\right] \cdot [i\ j\ k\ l]^T \quad , \tag{3}$$

where $a_i = bg_x$, $a_j = bg_y$, $a_k = zg_x/f$ and $a_l = zg_y/f$ and l_0 collects all the constant terms. The gradient of L, $\nabla L = \left[a_i\ a_j\ a_k\ a_l\right]^T$, contains the depth z only in the (k,l) derivatives. The only other unknown parameters, g_x and g_y, are present in both (i,j) and (k,l) and so can be cancelled by dividing a_k with a_i or a_l with a_j. Hence, the affine light field produces directly a depth estimate

$$z = bf\frac{a_k}{a_i} \quad \text{and/or} \quad z = bf\frac{a_l}{a_j} \quad . \tag{4}$$

Comparing Eq. 4 with stereo reconstruction, one finds, similarly, the baseline and focal length, while a_i/a_k and a_j/a_l do the role of disparities.

In order to use Eq. 4 to extract depth from a real scene, one has to estimate the values of $a_{(\cdot)}$, by calculating a locally affine approximation. This can be done by estimating the gradients in the EPI's, based on Sobel operators, as in [1]. Alternatively, in [5] the structure tensor is used, which involves derivative estimates in the four components of the light field combined with low pass filtering in the four dimensions, in order to attenuate high frequency noise enhanced by the derivative operations. We use the structure tensor formulation. When both a_i and a_j are not zero we output the mean of the two z estimates from Eq. 4. If just one value is not zero then the depth estimate is based just on that value.

4 Experiments and Results

In a first experiment a synthetic figure is created following the setup in Fig. 1. The camera parameters in our experiment are $b = 3 \times 10^{-4} m$ and $f = 200$, while the parameters of the scene are given by $\mathbf{g} = (1,0) m^{-1}$ and $z = 0.15m$, which theoretically results in a light field given by $L = 10^{-4} i + 7.5 \times 10^{-4} k$. From the resulting lightfield, the gradients can be extracted using the structure tensor as in [5] without the regularization step. Even in this simple setup, one has to contend with errors induced by quantization of the image signal, in our case 8 bits. Nonetheless, the reconstruction returned robust results of $z = 0.149 \pm 0.008\ m$.

In a second experiment we considered a more involved setting, a spherical hubcap on top of a plane with a gradient, as represented in Fig. 3. In this case the light field is not globally affine on the hubcap. The same reconstruction method was applied with the results illustrated in Fig. 4. Good results were obtained even on non globally affine light fields, since they are still locally affine, i.e. are well represented locally by a first order approximation. The mean of the absolute relative errors obtained was 1.49%.

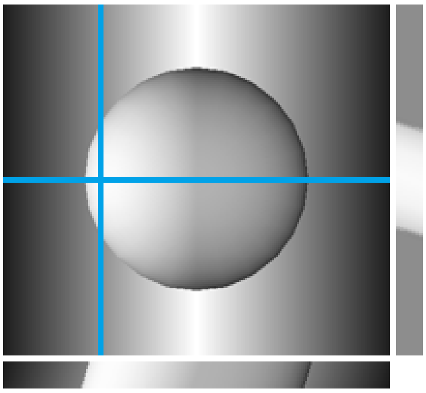

Figure 3: Example light field image to demonstrate depth reconstruction. Central viewpoint surrounded by two EPI's. The bottom and right EPI's originate from the horizontal and vertical lines, respectively.

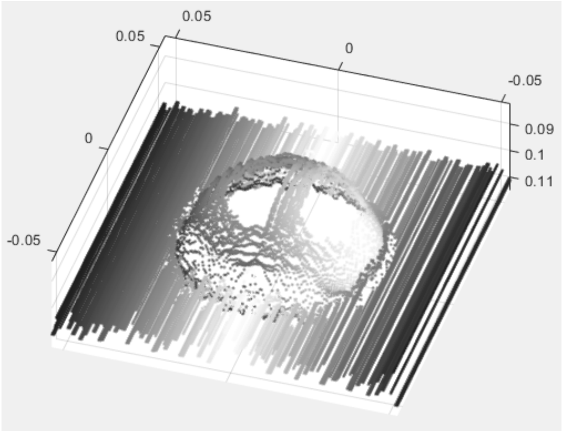

Figure 4: Reconstruction of the synthetic light field image. Depth values are measured with respect to the camera coordinates frame.

5 Conclusions

In this paper we have shown how a light field camera model and its produced images can be interpreted in familiar terms, so as to facilitate the reconstruction of the 3D objects captured. Furthermore, we introduced a minimal order light-field containing depth information which can be extracted by light-field analysis

References

[1] Don Dansereau and Len Bruton. Gradient-based depth estimation from 4d light fields. In *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on*, volume 3, pages III–549. IEEE, 2004.

[2] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1027–1034, 2013.

[3] Simao Graça Marto, Nuno Barroso Monteiro, Joao Pedro Barreto, and José António Gaspar. Structure from plenoptic imaging. In *IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*, volume 18, page 21, 2017.

[4] Ren Ng. *Digital light field photography*. PhD thesis, stanford university, 2006.

[5] Sven Wanner, Christoph Straehle, and Bastian Goldluecke. Globally consistent multi-label assignment on the ray space of 4d light fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1011–1018, 2013.