# Plenoptic Video Geometry

Jan Neumann and Cornelia Fermüller
Center for Automation Research
University of Maryland
College Park, MD 20742-3275, USA
{jneumann,fer}@cfar.umd.edu

**Keywords:** Multi-view Geometry, Spatio-temporal Image Analysis, Camera Design, Structure from Motion, Polydioptric Cameras

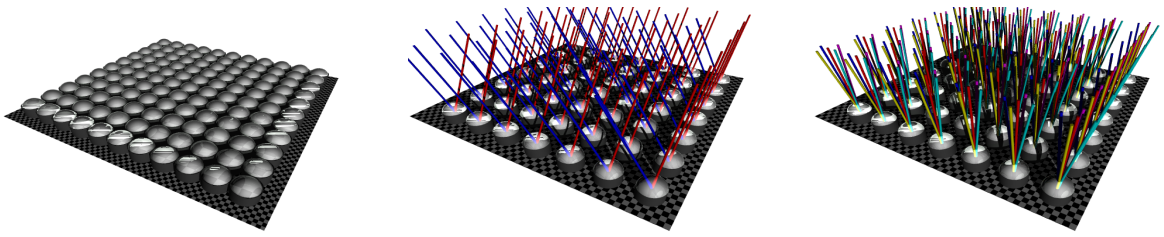
## Abstract

More and more processing of visual information is nowadays done by computers, but the images captured by conventional cameras are still based on the pinhole principle inspired by our own eyes. This principle though is not necessarily the optimal image formation principle for automated processing of visual information. Each camera samples the space of light rays according to some pattern. If we understand the structure of the space formed by the light rays passing through a volume of space, we can determine the camera, or in other words the sampling pattern of light rays, that is optimal with regard to a given task. In this work we analyze the differential structure of the space of time-varying light rays described by the plenoptic function and use this analysis to relate the rigid motion of an imaging device to the derivatives of the plenoptic function. The results can be used to define a hierarchy of camera models with respect to the structure from motion problem and formulate a linear, scene-independent estimation problem for the rigid motion of the sensor purely in terms of the captured images.

## 1 Introduction

When we think about vision, we usually think of interpreting the images taken by (two) eyes such as our own - that is, images acquired by camera-type eyes based on the pinhole principle. This advantageous because because it enables an easy interpretation of the visual information by a human observer. Despite the fact that many image interpretation tasks are automated nowadays, most commercially available cameras are still based on the same pinhole principle. One considers a point in space and the light rays passing through that point. Then the rays are cut with an imaging surface, and a subset of them forms an image. This image captures a view of the world that looks very similar to the view we capture with our own eyes, but this might not be optimal image to solve a given task. Of course these are not the only types of eyes that exist; the biological world reveals a large variety of eye designs. It has been estimated that eyes have evolved no fewer than forty times, independently, in diverse parts of the animal kingdom [8], and these eye designs, and therefore the images they capture, are highly adapted to the tasks the animal has to perform. This suggests that we should not just focus our efforts on designing algorithms that optimally process a given visual input, but also optimize the design of the imaging sensor with regard to the task at hand, so the subsequent processing of the visual information is facilitated.

This focus on sensor design has already begun. Technological advances make it possible to construct integrated imaging devices using electronic and mechanical microassembly, micro-optics, and advanced data buses, not only of the kind that exists in nature such as log-polar retinas [5], but also of many other kinds such as catadioptric cameras [15]. Nevertheless, a general framework to relate the design of an imaging sensor to its usefulness for a



(a) (b) (c)

Figure 1: Design of a Polydioptric Camera (a) capturing Parallel Rays (b) and simultaneously capturing a Pencil of Rays (c).

given task is still missing.

In this work we will present such a framework by studying the relationship between the subset of light rays captured by a generalized camera and its performance with regard to the task of egomotion estimation. To analyze these general cameras, we will study the most complete visual representation of a scene, namely the plenoptic function as it changes differentially over time [1]. In free space the plenoptic function reduces to the 5D space of time-varying light rays. Since any imaging device captures a subset of this space, the choice of the subset determines how well a task can be performed. We will use our study of the differential structure of the time-varying plenoptic function captured by a rigidly moving imaging sensor to analyze how this structure is related to the rigid motion of the sensor.

The idea of studying the space of light rays was already mentioned by Leonardo da Vinci [20] and further studied in the context of photometry at the beginning of the 20th century (for an overview see [14]). In computer graphics and computer vision recent work uses non-perspective subsets of the plenoptic function to represent visual information to be used for image-based rendering. Some examples are light fields [13] and lumigraphs [9], multiple centers of projection images [19] which have been used in cell animation already for quite some time [23], or multi-perspective panoramas [18]. Non-perspective images have also been used by several people to reconstruct the observed scene from video sequences (for some examples see [3, 22, 6]).

In general terms, a camera is a mechanism that forms images by focusing light onto a light sensitive surface (retina, film, CCD array, etc.). Different cameras are obtained by varying three elements: (1) the geometry of the surface, (2) the geometric distribution and optical properties of the photoreceptors, and (3) the way light is collected and projected onto the surface (single or multiple lenses, or tubes as in compound eyes).

A theoretical model for a camera that captures the plenoptic function in some part of the space is a surface $S$ that has at every point a pinhole camera (for other possible parameterizations for generalized cameras see [21, 10, 17]). We call the actual implementation of this theoretical concept a *polydioptric* camera. A "plenoptic camera" has been proposed in [2], but since no physical device can capture the true time-varying plenoptic function, we prefer the term polydioptric to emphasize the difference between the theoretical concept and the implementation. With a polydioptric camera we observe points in the scene in view from many different viewpoints (theoretically, from every point on $S$) and thus we capture many rays emanating from that point.

A polydioptric camera can be implemented by arranging ordinary cameras very close to each other (Figs. 1b and 1c). This camera has an additional property arising from the proximity of the individual cameras: it can form a very large number of orthographic images, in addition to the perspective ones.

Indeed, consider a direction $\boldsymbol{r}$ in space and then consider in each individual camera the captured ray parallel to $\boldsymbol{r}$. All these rays together, one from each camera, form an image with rays that are parallel. For different directions $\boldsymbol{r}$ different orthographic image are formed. For example, Fig. 1b shows that we can select one appropriate pixel in each camera to form an orthographic image that looks to one side (blue rays pointing to the left) or another (red

rays pointing to the right). Fig. 1c shows all the captured rays, thus illustrating that each individual camera collects conventional pinhole images.

In conclusion, a polydioptric camera has the unique property that it captures, simultaneously, a large number of perspective and orthographic images (projections). We will demonstrate later that this capability enables us to solve the structure from motion problem using a linear, scene-independent equation.

## 2 Plenoptic Video Geometry

Let the scene surrounding the image sensor be modeled by the signed distance function $f(\boldsymbol{x};t) : \mathbb{R}^3 \times \mathbb{R}_+ \rightarrow \mathbb{R}$ where $f(\boldsymbol{x};t) > 0$ if the world point $\boldsymbol{x}$ is not occupied by a scene object, and $f(\boldsymbol{x};t) \leq 0$ if $\boldsymbol{x}$ lies inside a scene object.

At each location $\boldsymbol{x}$ in free space ($f(\boldsymbol{x};t) > 0$), the radiance, that is the light intensity or color observed at $\boldsymbol{x}$ from a given direction $\boldsymbol{r}$ at time $t$, is measured by the plenoptic function $L(\boldsymbol{x};\boldsymbol{r};t)$; $L : \mathbb{R}^3 \times \mathbb{S}^2 \times \mathbb{R}_+ \rightarrow \mathbb{R}^d$, where $d = 1$ for intensity, $d = 3$ for color images, and $\mathbb{S}^2$ is the unit sphere of directions in $\mathbb{R}^3$.

The image irradiance that is recorded by an imaging device is proportional to the scene radiance [11], therefore, we assume that the intensity recorded by an imaging sensor at position $\boldsymbol{x}$ and time $t$ pointing in direction $\boldsymbol{r}$ is equal to the plenoptic function $L(\boldsymbol{x};\boldsymbol{r};t)$.

Since a transparent medium such as air does not change the color of the light, the radiance along the view direction $\boldsymbol{r}$ is constant:

$$L(\boldsymbol{x};\boldsymbol{r};t) = L(\boldsymbol{x}+\lambda\boldsymbol{r};\boldsymbol{r};t)\ \forall\lambda : f(\boldsymbol{x}+\lambda\boldsymbol{r}) > 0$$

which implies

$$\nabla_{\boldsymbol{x}} L^{\mathrm{T}}\boldsymbol{r} = \nabla_{\boldsymbol{r}} L^{\mathrm{T}}\boldsymbol{r} = 0\ \forall\boldsymbol{x} \in \mathbb{R}^3 \text{ with } f(\boldsymbol{x};t) > 0$$

where $\nabla_{\boldsymbol{x}} L$ and $\nabla_{\boldsymbol{r}} L$ are the partial derivatives of $L$ with respect to $\boldsymbol{x}$ and $\boldsymbol{r}$. Therefore, the plenoptic function in free space reduces to five dimensions – the time-varying space of directed lines for which many representations have been presented (for an overview see Camahort and Fussel [4]).

Let us assume that the albedo of every scene point is constant over time and that we observe a static world under constant illumination. In this case, the radiance of a light ray does not change over time which implies that the total time derivative of the plenoptic function vanishes:

$$\frac{d}{dt} L(\boldsymbol{x};\boldsymbol{r};t) = 0.$$

If at the intersection point $\boldsymbol{y} \in \mathbb{R}^3$ of the ray $\phi$ ($\phi(\lambda) = \boldsymbol{x} + \lambda\boldsymbol{r}$) with the scene surface ($f(\boldsymbol{y}) = 0$) the albedo $\rho(\boldsymbol{y};t)$ is continuously varying and no occlusion boundaries are present ($\boldsymbol{r}^{\mathrm{T}}\nabla f(\boldsymbol{y},t) \neq 0$), then we can develop the plenoptic function $L$ in the neighbourhood of $(\boldsymbol{x};\boldsymbol{r};t)$ into a Taylor series ($L_t$ is an abbreviation for $\partial L/\partial t$):

$$L(\boldsymbol{x}+d\boldsymbol{x};\boldsymbol{r}+d\boldsymbol{r};t+dt) = L(\boldsymbol{x};\boldsymbol{r};t) \qquad (1)$$
$$+L_t dt + \nabla_{\boldsymbol{x}} L^{\mathrm{T}} d\boldsymbol{x} + \nabla_{\boldsymbol{r}} L^{\mathrm{T}} d\boldsymbol{r} + \mathcal{O}(\|d\boldsymbol{r}, d\boldsymbol{x}, dt\|^2).$$

This expression now relates a local change in view ray position and direction to the first-order differential brightness structure of the plenoptic function.

We define the *plenoptic ray flow* as the difference in position and orientation of rays between the two rays that are captured by the same imaging element at two consecutive time instants. This allows us to use the spatio-temporal brightness derivatives of the light rays captured by an imaging device to constrain the plenoptic ray flow by generalizing the well-known *Image Brightness Constancy Constraint* to the *Plenoptic Brightness Constancy Constraint*:

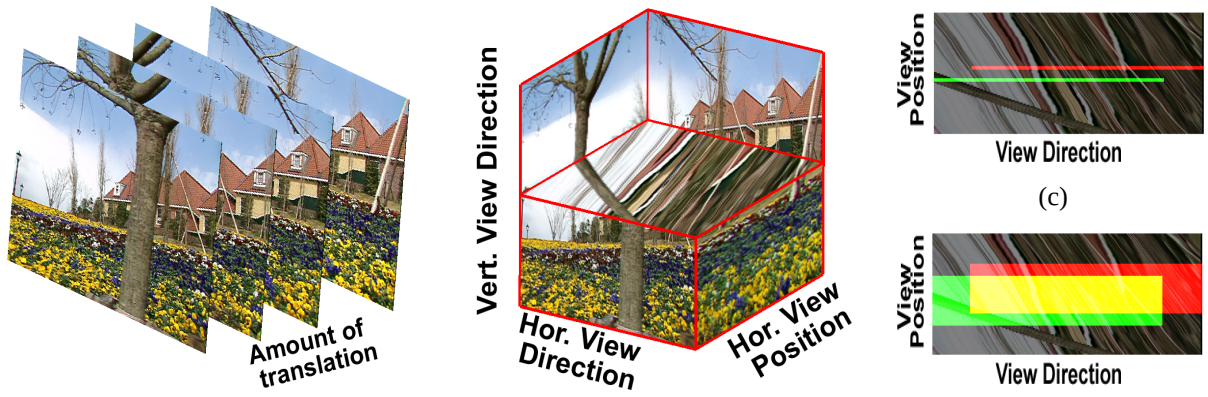
$$\frac{d}{dt} L(\boldsymbol{r};\boldsymbol{x};t) = L_t + \nabla_{\boldsymbol{r}} L^{\mathrm{T}} \frac{d\boldsymbol{r}}{dt} + \nabla_{\boldsymbol{x}} L^{\mathrm{T}} \frac{d\boldsymbol{x}}{dt} = 0. \quad (2)$$

## 3 Plenoptic Motion Equations

The set of imaging elements that make up a camera each capture the radiance at a given position coming from a given direction. If the camera undergoes a rigid motion, then we can express this motion by a rigid coordinate transformation of the ambient space of light rays. This rigid transformation, parameterized by the rotation matrix $R$ and a translation vector $\boldsymbol{t}$, results in the *exact* identity (see the example in Fig. 2)

$$L(\boldsymbol{x};\boldsymbol{r};t) = L(R\boldsymbol{x}+t;R\boldsymbol{r};t+1)$$

since the rigid motion maps the time-invariant space of light rays upon itself. Thus, the problem of estimating the




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

Figure 2: (a) Sequence of images captured by a horizontally translating camera. (b) Epipolar plane image volume formed by the image sequence where the top half of the volume has been cut away to show how a row of the image changes when the camera translates. (c) A row of an image taken by a pinhole camera at two time instants (red and green) corresponds to two non-overlapping horizontal line segments in the epipolar plane image, while in (d) the collection of corresponding "rows" of a polydioptric camera at two time instants corresponds to two rectangular regions of the epipolar image that do overlap (yellow region). This overlap enables us to estimate the rigid motion of the camera purely based on the visual information recorded.

rigid motion of a sensor has become an image registration problem that is *independent of the scene* and only depends on the rigid motion parameters!

We assume that the imaging sensor can capture images at a rate that allows us to use the small angle approximation for the rotation matrix $R \approx I + [\boldsymbol{\omega}]_x$ where $[\boldsymbol{\omega}]_x$ is a skew-symmetric matrix parameterized by the axis of the instantaneous rotation $\boldsymbol{\omega}$. Now we can define the plenoptic ray flow for the ray captured by the imaging element located at location $\boldsymbol{x}$ and looking in direction $\boldsymbol{r}$ as

$$\frac{d\boldsymbol{r}}{dt} = \boldsymbol{\omega} \times \boldsymbol{r} \text{ and } \frac{d\boldsymbol{x}}{dt} = \boldsymbol{\omega} \times \boldsymbol{x} + \boldsymbol{t} \tag{3}$$

where $\boldsymbol{t}$ is the instantaneous translation. In stark contrast to the projection of the rigid motion flow on the imaging surface of a conventional pinhole camera which depends on an infinite number of parameters due to its dependence on the scene depth structure, we see that the plenoptic ray flow is completely specified by the six rigid motion parameters. This regular global structure of the rigid plenoptic ray flow makes the estimation of the rigid motion parameters very well-posed.

Combining Eqs. 2 and 3 leads to the *plenoptic motion constraint*

$$\begin{aligned} -L_t &= \nabla_{\boldsymbol{x}} L \cdot (\boldsymbol{\omega} \times \boldsymbol{x} + \boldsymbol{t}) + \nabla_{\boldsymbol{r}} L \cdot (\boldsymbol{\omega} \times \boldsymbol{r}) \\ &= \nabla_{\boldsymbol{x}} L \cdot \boldsymbol{t} + (\boldsymbol{x} \times \nabla_{\boldsymbol{x}} L + \boldsymbol{r} \times \nabla_{\boldsymbol{r}} L) \cdot \boldsymbol{\omega} \end{aligned} \tag{4}$$

which is a linear, scene-independent constraint in the motion parameters and the plenoptic partial derivatives.

This formalism can also be applied if we observe a rigidly moving object with a set of static cameras. In this case, we attach the world coordinate system to the moving object and we can relate the relative motion of the image sensors with respect to the object to the spatio-temporal derivatives of the light rays that leave the object. It is to note though, that we need to account for illumination effects due to the relative motion between object and light sources.

It is important to realize that the derivatives $\nabla_{\boldsymbol{r}} L$ and $\nabla_{\boldsymbol{x}} L$ can be obtained from the image information captured by a polydioptric camera. Recall that a polydioptric camera can be envisioned as a surface where every point corresponds to a pinhole camera. $\nabla_{\boldsymbol{r}} L$, the plenoptic derivative with respect to direction, is the derivative

with respect to the image coordinates that one finds in a traditional pinhole camera. One keeps the position and time constant and changes direction (Fig. 1c).

The second plenoptic derivative, $\nabla_{\boldsymbol{x}} L$, is obtained by keeping the direction of the ray constant and changing the position along the surface (Fig. 1b). Thus, one captures the change of intensity between parallel rays. This is similar to computing the derivatives in an orthographic camera. In Section 1 we mentioned that a polydioptric camera captures perspective and orthographic images. $\nabla_{\boldsymbol{r}} L$ is found from the perspective images and $\nabla_{\boldsymbol{x}} L$ from the orthographic images.

The ability to compute all the plenoptic derivatives depends on the ability to capture light at multiple viewpoints coming from multiple directions. This corresponds to the ability to incorporate stereo information into motion estimation, since multiple rays observe the same part of the world. For single-viewpoint cameras this is inherently impossible, and thus it necessitates nonlinear estimation over both structure and motion to compensate for this lack of multi-view (or equivalently depth) information.

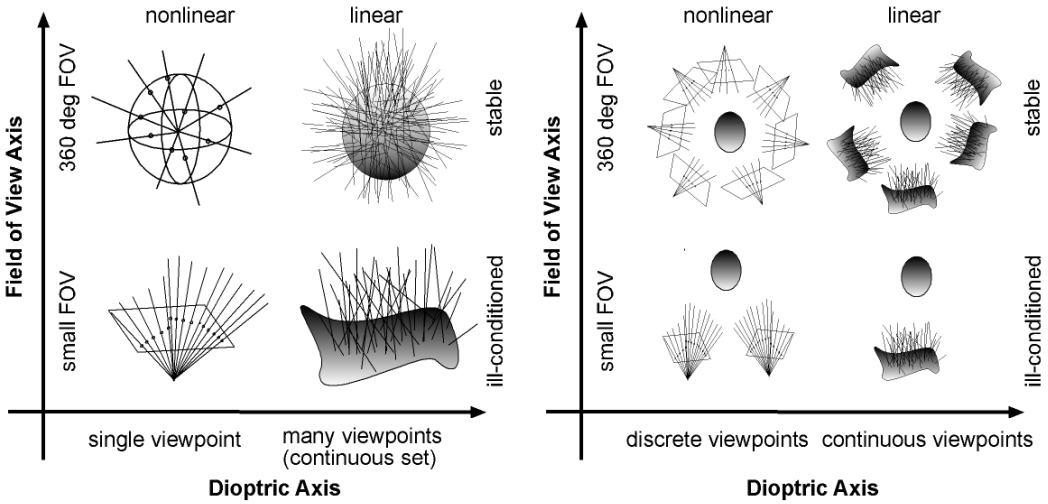
To illustrate this idea further, we translate a camera along the horizontal image axis and stack the images to form an image volume (Figs. 2a-2b). Due to the horizontal translation scene points always project into the same row in each of the images. Such an image volume is known as an epipolar image volume [3] since corresponding rows lie all in the same epipolar plane. A horizontal slice through the image volume (Figs. 2c-2d) is called an epipolar plane image and illustrates the structure of the visual space in dependence on view position and direction. A row of an image taken by a pinhole camera corresponds to a horizontal line segment in the epipolar plane image (Fig. 2c), while the collection of corresponding "rows" of a polydioptric camera captures a rectangular area of the epipolar image (Fig. 2d). Here we assumed that the viewpoint axis of the polydioptric camera is aligned with the direction of translation used to define the epipolar image volume. If this is not the case, we can warp the images to align the optical centers. We see that a camera rotation around an axis perpendicular to the epipolar image plane corresponds to a horizontal shift in the epipolar image, while a translation of the camera parallel to an image row, causes a vertical shift. These shifts can be different for each pixel depending on the rigid motion of the camera. If we want to recover this rigid transformation based on the images captured, we see in (Fig. 2c) that when using a pinhole camera, we have to match two non-overlapping subsets of the visual space (shown in green and red). At each time a pinhole camera captures by definition only the view from a single viewpoint. Therefore, it is necessary for an accurate recovery of the rigid motion that we estimate the scene structure, since the correspondence between pixels in image rows taken from different view points depends on the depth of the scene.

In contrast, we see in (Fig. 2d) that for a polydioptric camera the matching can be based purely on the captured image information, since we have a "natural" brightness constancy in the region of overlap (yellow) since we match a light ray with itself. In this case the correspondence of pixels (light rays) depends only on the motion of the camera, not on the depth of the scene, thus enabling us to estimate the rigid motion of the camera in a completely scene-independent manner.

In previous work [16], we developed a hierarchy of camera design with regard to the linearity of the estimation and the stability due to the field of view (see Fig. 3). We see in the figure that although the estimation of structure and motion for a single-viewpoint spherical camera is stable and robust, it is still non-linear, and the algorithms which give the most accurate results are search techniques, and thus rather elaborate. A spherical polydioptric camera combines the stability of full field of view motion estimation with the linearity and scene independence of the problem, and is therefore the camera of choice to solve the structure from motion problem.

# 4 Light Field Video Geometry

Due to the difficulties involved when using signal processing operators in a spherical coordinate system, we will choose the two-plane parameterization that was used by [9, 13] to represent the space of light rays. All the lines passing through some space of interest can be parameterized by surrounding this space (that could contain either a camera or an object) with two nested cubes and then recording the intersection of the light rays entering the camera or leaving the object with the planar faces of the two cubes. We only describe the parameterization of the rays passing through one pair of faces, the extension to the other pairs is straight forward. Without loss of generality




(a) Static World - Moving Cameras (b) Moving Object - Static Cameras

Figure 3: Hierarchy of Cameras. We classify the different camera models according to the field of view (FOV) and the number and proximity of the different viewpoints that are captured (Dioptric Axis). This in turn determines if structure from motion estimation is a well-posed or an ill-posed problem, and if the motion parameters are related linearly or non-linearly to the image measurements. The camera models are clockwise from the lower left: small FOV pinhole camera, spherical pinhole camera, spherical polydioptric camera, and small FOV polydioptric camera.

we choose both planes to be perpendicular to the z-axis and seperated by a distance of $f$. We denote one plane as *focal plane* $\Pi_f$ indexed by coordinates $(x, y)$ and the other plane as *image plane* $\Pi_i$ indexed by $(u, v)$, where $(u, v)$ is defined in a local coordinate sytem with respect to $(x, y)$ (see Fig. 4a). Both $(x, y)$ and $(u, v)$ are aligned with the $(X, Y)$-axes of the world coordinates and $\Pi_f$ is at a distance of $Z_\Pi$ from the origin of the world coordinate system.

This enables us now to parameterize the light rays that pass through both planes at any time $t$ using the tupels $(x, y, u, v, t)$ and we can record their intensity in the time-varying light field $L(x, y, u, v, t)$. For fixed location $(x, y)$ in the focal plane, $L(x, y, \cdot, \cdot, t)$ corresponds to the image captured by a perspective camera. If instead we fix the view direction $(u, v)$, we capture an orthographic image $L(\cdot, \cdot, u, v, t)$ of the scene ($\cdot$ denotes the varying parameters).

Using this light field parameterization we can rewrite the plenoptic motion equation (Eq. 4) by setting $\boldsymbol{x} = [x, y, Z_\Pi]^{\mathrm{T}}$ and $\boldsymbol{r} = \frac{[u,v,f]^{\mathrm{T}}}{\|[u,v,f]^{\mathrm{T}}\|}$. To be able to use this equation we need to convert the spatial partial derivatives of the light field $L_x = \partial L/\partial x, \ldots, L_v = \partial L/\partial v$ into the three-dimensional plenoptic derivatives $\nabla_{\boldsymbol{x}} L$ and $\nabla_{\boldsymbol{r}} L$.

We consider the projections of $\nabla_{\boldsymbol{x}} L$ and $\nabla_{\boldsymbol{r}}$ on three directions $\boldsymbol{c}_x$,$\boldsymbol{c}_r$, and $\boldsymbol{r}$ to obtain the following linear system which we solve for $\nabla_{\boldsymbol{x}} L$ and $\nabla_{\boldsymbol{r}} L$ ($\boldsymbol{c}_x = [1, 0, 0]^{\mathrm{T}}$, $\boldsymbol{c}_y = [0, 1, 0]^{\mathrm{T}}$, and $n_{\boldsymbol{r}} = \|[u, v, f]^{\mathrm{T}}\|$):

$$\begin{pmatrix} \boldsymbol{c}_x^{\mathrm{T}} \\ \boldsymbol{c}_y^{\mathrm{T}} \\ \boldsymbol{r}^{\mathrm{T}} \end{pmatrix} [\nabla_{\boldsymbol{x}} L, \nabla_{\boldsymbol{r}} L] = \begin{pmatrix} L_x & n_{\boldsymbol{r}} L_u \\ L_y & n_{\boldsymbol{r}} L_v \\ 0 & 0 \end{pmatrix}.$$

This results in the following expressions for the plenoptic derivatives

$$\nabla_{\boldsymbol{x}} L = [L_x, L_y, -\frac{u}{f} L_x - \frac{v}{f} L_y]^{\mathrm{T}}$$
$$\nabla_{\boldsymbol{r}} L = n_{\boldsymbol{r}} [L_u, L_v, -\frac{u}{f} L_u - \frac{v}{f} L_v]^{\mathrm{T}}$$

If we now plug these expressions into Eq.4, we can define the plenoptic motion constraint for the ray indexed by $(x, y, u, v, t)$ as ($[\cdot; \cdot]$ denotes the vertical stacking of

vectors):

$$\begin{aligned}
-L_t &= \nabla_{\boldsymbol{x}} L \cdot \boldsymbol{t} + (\boldsymbol{x} \times \nabla_{\boldsymbol{x}} L + \boldsymbol{r} \times \nabla_{\boldsymbol{r}} L) \cdot \boldsymbol{\omega} \\
&= [L_x, L_y, -\frac{u}{f} L_x - \frac{v}{f} L_y]\boldsymbol{t} \\
&\quad - [L_x, L_y, -\frac{u}{f} L_x - \frac{v}{f} L_y]([x, y, Z_\Pi]^{\mathrm{T}} \times \boldsymbol{\omega}) \\
&\quad - [L_u, L_v, -\frac{u}{f} L_u - \frac{v}{f} L_v]([u, v, f]^{\mathrm{T}} \times \boldsymbol{\omega}) \\
&= [L_x, L_y, L_u, L_v][M_t, M_\omega][\boldsymbol{t}; \boldsymbol{\omega}] \qquad (5)
\end{aligned}$$

where

$$M_t = \begin{pmatrix} 1 & 0 & -\frac{u}{f} \\ 0 & 1 & -\frac{v}{f} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$M_\omega = \begin{pmatrix} -\frac{uy}{f} & \frac{ux}{f} + Z_\Pi & -y \\ -(\frac{vy}{f} + Z_\Pi) & \frac{vx}{f} & x \\ -\frac{uv}{f} & \frac{u^2}{f} + f & -v \\ -(\frac{v^2}{f} + f) & \frac{vu}{f} & u \end{pmatrix}$$

By combining the constraints across the light field, we can form a highly over-determined linear system and solve for the rigid motion parameters.

As said before, the light field derivatives $L_x, \ldots, L_t$ can be obtained directly from the image information captured by a polydioptric camera. To convert the image information captured by this collection of pinhole cameras into a light field, for each camera we simply have to intersect the rays from its optical center through each pixel with the two planes $\Pi_f$ and $\Pi_i$ and set the corresponding light field value to the pixel intensity. Since our measurements are in general scattered, we have to use appropriate interpolation schemes to compute a continuous light field function (for an example the push-pull scheme in [9]). The light field derivatives can then easily be computed by applying standard image derivative operators to the continuous light field. The plenoptic motion constraint is extended to the other faces of the nested cube by premultiplying $\boldsymbol{t}$ and $\boldsymbol{\omega}$ with the appropriate rotation matrices to rotate the motion vectors into local light field coordinates.

The relationship between the structure from motion formulation for conventional single viewpoint cameras and the formulation for polydioptric cameras can easily be established. If we consider a 2-D slice of the light field corresponding to the indices $(x, u)$ that form an epipolar image [3] and assume that the surface has slowly varying reflectance, we can apply a simple triangulation argument to get the following identity ($du$ is the change in view direction and $dx$ the corresponding change in view position as illustrated in Fig. 4b).

$$\begin{aligned}
&L(x, y, u + du, v + dv, t) \\
=&L(x, y, u + \frac{f}{Z} dx, v + \frac{f}{Z} dy, t) \qquad (6) \\
=&L(x + dx, y + dy, u, v, t)
\end{aligned}$$

If we now compute the directional derivative of the light field with respect to the $u$-axis (the derivation with respect to the $v$-axis is analog), we get

$$\begin{aligned}
L_u = \frac{dL}{du} &= \lim_{\Delta u \to 0} \frac{L(x, u + \Delta u, t) - L(x, u, t)}{\Delta u} \\
&= \lim_{\Delta u \to 0} \frac{L(x + \frac{Z}{f} \Delta u, u, t) - L(x, u, t)}{\Delta u} \\
&= \lim_{\Delta u \to 0} \frac{L(x + \Delta u, u, t) - L(x, u, t)}{\frac{f}{Z} \Delta u} \\
&= \frac{Z}{f} \frac{dL}{dx} = \frac{Z}{f} L_x
\end{aligned}$$

and see that depth is encoded as the ratio between the positional and directional derivatives where $Z$ denotes the depth of the scene point imaged by the ray with index tupel $(x, y, u, v)$.

From Eq. 6, we can conclude now that we can replace any flow in the view direction variables $(u, v)$ that is inversely proportional to the depth of the scene, e.g. the projection of the translational flow, by flow in the view points $(x, y)$ that is independent of the scene. This enables us to formulate the structure from motion problem independent of the scene in view as a linear problem. We also see that the depth $Z$ is proportional to the ratios between the positional and directional derivatives $Z/f = L_u/L_x = L_v/L_y$. This is identical to the relation between depth and differential image measurements used in differential stereo and in the epipolar-plane image analysis [3]. Thus, we can interpret plenoptic motion estimation as the integration of differential motion estimation with differential stereo.

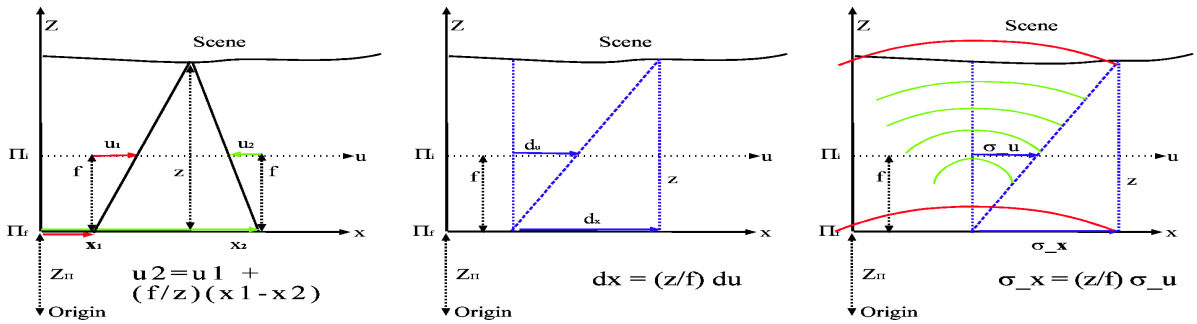



Figure 4: (a) Light Field Parameterization (here shown only for the light field slice spanned by axes $x$ and $u$) (b) The ratio between the magnitude of perspective and orthographic derivatives is linear in the depth of scene. (c) The scale at which the perspective and orthographic derivatives match depends on the depth (matching scale in red)

### 4.1 Scale Dependence of the Plenoptic Derivatives

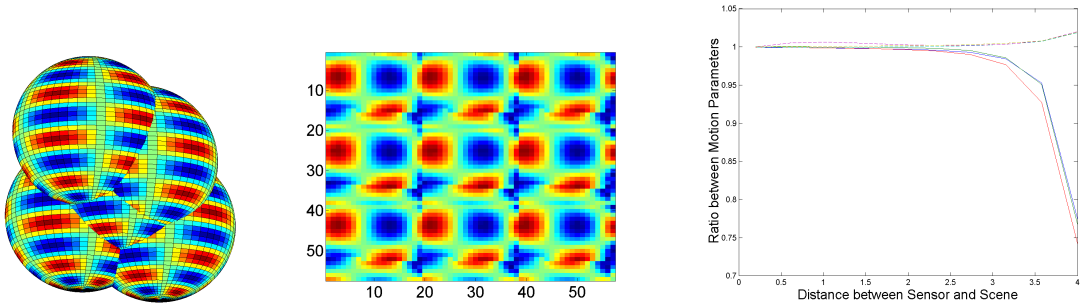
One issue we will not address in this paper but that needs to be considered is how densely the plenoptic space needs to be sampled by a polydioptric camera to be able to compute accurate structure and motion information. A similar problem has been studied in computer graphics under the name plenoptic sampling [7] with the aim of improving view interpolation in image-based rendering. The sampling density necessary to avoid aliasing depends of course on the brightness profile and depth structure of the scene, thus the choice of the correct smoothing filter depends very much on prior knowledge. The question how to design this optimal smoothing filter for motion estimation is beyond the scope of this paper.

The accuracy of the linearization of the time-varying light field (Eq. 1) also depends on the compatibility of the plenoptic derivatives. This means that the estimation of the perspective, orthographic, and temporal plenoptic derivatives needs to be based upon similar subsets of the scene radiance.

Notice that if we combine information from neighboring measurements in directional space at a fixed position, we integrate radiance information over a region of the scene surface whose area scales with the distance from the sensor to the scene. In contrast, if we combine information over neighboring measurements in positional space for a fixed direction, we integrate information over a region of the scene surface whose area is independent of the depth (illustrated in Fig.4c). Unless the brightness structure of the scene has enough similarity across scales (e.g., if the local scene radiance changes linearly on the scene surface), so that the derivative computation is invariant to our choice of filter size, we have to make sure when we compute the plenoptic derivatives with respect to time, direction, and position that the domains of integration of our derivative filters relative to the scene are as similar as possible. One way to adjust the integration domains of the filters would be to compute the temporal, directional and positional derivatives at many scales and use Eq. 7 as a constraint to find the best relative shift in scale space between the positional, directional, and temporal derivatives.

## 5 Experiments

To examine the performance of an algorithm using the plenoptic motion constraint, we did experiments with synthetic data. We distributed spheres, textured with a smoothly varying pattern, randomly in the scene so that they filled the horizon of the camera (see Fig. 5a). We then computed the light fields for all the faces of the nested cube surrounding the camera through raytracing, computed the derivatives, stacked the linear equations (Eq. 5) to form a linear system, and solved for the motion parameters. Even using derivatives only at one scale, we




(a) (b) (c)

Figure 5: (a) Subset of an Example Scene, (b) the corresponding light field (c) Accuracy of Plenoptic Motion Estimation. The plot shows the ratio of the true and estimated motion parameters (vertical axis) in dependence of the distance between the sensor surface and the scene (horizontal axis) for $f = 60$ and spheres of unit radius.

found that the motion is recovered very accurately as seen in Fig. 5c. As long as the relative scales of the derivatives were similar enough (scene not too far away) the error in the motion parameters varied between 1% and 3%. This accurate egomotion estimate could then be used to compute depth from differential measurements using the following four formulas ($M_x^i$ denotes the $i$th row of the coefficient matrix $M_t$ or $M_\omega$ in Eq. 5):

$$\begin{aligned} z &= \frac{L_u}{fL_x} = \frac{L_v}{fL_y} \\ &= -\frac{[L_u M_t^1 + L_v M_t^2]\boldsymbol{t} + [L_u M_\omega^1 + L_v M_\omega^2]\boldsymbol{\omega}}{f(L_t + [L_u M_\omega^3 + L_v M_\omega^4]\boldsymbol{\omega})} \\ &= -\frac{L_t + [L_x M_t^1 + L_y M_t^2]\boldsymbol{t} + [L_x M_\omega^1 + L_y M_\omega^2]\boldsymbol{\omega}}{f[L_x M_\omega^3 + L_y M_\omega^4]\boldsymbol{\omega}} \end{aligned} \tag{7}$$

where the first two equations describe the well-know constraints of differential stereo, while the last two equations have been used in the community to compute depth maps based on structure from motion constraints.

The analysis suggests the following plenoptic structure from motion algorithm. Using the proposed plenoptic motion framework, one can envision a feedback loop algorithm, where we use all the plenoptic derivatives to compute an estimate of the camera motion using Eq. 4. Since we are solving a linear system, the computation of the motion parameters is fast and we do not have any convergence issues as in the nonlinear methods necessary for single-viewpoint cameras. Then we can use the recovered motion together with the plenoptic derivatives to compute a scene depth estimate. If the four estimates in Eq. 7 do not match, we adapt the integration domains of the temporal, directional and positional derivative filters until we compute consistent depth and motion estimates. This is repeated for each frame of the input video, while simultaneously we use the computed motion trajectory to integrate and refine the instantaneous depth maps in a large-baseline stereo optimization to construct accurate three-dimensional descriptions or image-based representations (e.g. [12]) of the scene.

# 6 Conclusion

According to ancient Greek mythology Argus, the hundred-eyed guardian of Hera, the goddess of Olympus, alone defeated a whole army of Cyclopes, one-eyed giants. The mythological power of many eyes became real in this paper, which proposed a mathematical analysis of the differential structure of the space of light rays. We introduced a framework to systematically study the relationship between the shape of an imaging sensor and the task performance of the entity using this sensor. In this paper we focused on the relation between the local differential structure of the time-varying plenoptic function and the ego-motion estimation of an imaging sensor. Of course many more tasks are imaginable (for example shape and 3D motion estimation from non-

perspective imagery). The application of this framework to the structure from motion problem resulted in a novel linear and scene-independent constraint between the differential structure of the plenoptic function and motion parameters describing the instantaneous motion of the sensor which were used to define guidelines for an optimal structure from motion camera design.

# References


[1] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In M. Landy and J. A. Movshon, editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press, Cambridge, MA, 1991.

[2] E. H. Adelson and J. Y. A. Wang. Single lens stereo with a plenoptic camera. *IEEE Trans. PAMI*, 14:99–106, 1992.

[3] 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*, 1:7–55, 1987.

[4] E. Camahort and D. Fussell. A geometric study of light field representations. Technical Report TR99-35, Dept. of Computer Sciences, The University of Texas at Austin, 1999.

[5] C. Capurro, F. Panerai, and G. Sandini. Vergence and tracking fusing log-polar images. In *Proc. International Conference on Pattern Recognition*, 1996.

[6] J. Chai and H. Shum. Parallel projections for stereo reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 493 –500, 2000.

[7] J. Chai, X. Tong, and H. Shum. Plenoptic sampling. In *Proc. of ACM SIGGRAPH*, pages 307–318, 2000.

[8] R. Dawkins. *Climbing Mount Improbable*. Norton, New York, 1996.

[9] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. In *Proc. of ACM SIGGRAPH*, pages 43–54, 1996.

[10] M. D. Grossberg and S. K. Nayar. A general imaging model and a method for finding its parameters. In *Proc. International Conference on Computer Vision*, pages 108–115, 2001.

[11] B. K. P. Horn. *Robot Vision*. McGraw Hill, New York, 1986.

[12] R. Koch, M. Pollefeys, B. Heigl, L. VanGool, and H. Niemann. Calibration of hand-held camera sequences for plenoptic modeling. In *Proc. Int. Conf. Computer Vision*, pages 585–591, 1999.

[13] M. Levoy and P. Hanrahan. Light field rendering. In *Proc. of ACM SIGGRAPH*, pages 161–170, 1996.

[14] P. Moon and D.E. Spencer. *The Photic Field*. MIT Press, Cambridge, 1981.

[15] S. Nayar. Catadioptric omnidirectional camera. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 482–488, 1997.

[16] J. Neumann, C. Fermüller, and Y. Aloimonos. Eyes from eyes: New cameras for structure from motion. In *IEEE Workshop on Omnidirectional Vision 2002 (in conjunction with ECCV 2002)*, pages 19–26, 2002.

[17] T. Pajdla. Stereo with oblique cameras. *International Journal of Computer Vision*, 47(1/2/3):161–170, 2002.

[18] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *CVPR'97*, pages 338–343, June 1997.

[19] P. Rademacher and G. Bishop. Multiple-center-of-projection images. *Proc. of ACM SIGGRAPH*, pages 199–206, 1998.

[20] J. P. Richter, editor. *The Notebooks of Leonardo da Vinci*, volume 1, p.39. Dover, New York, 1970.

[21] S. Seitz. The space of all stereo images. In *Proc. International Conference on Computer Vision*, 2001.

[22] H.Y. Shum, A. Kalai, and S. M. Seitz. Omnivergent stereo. In *Proc. International Conference on Computer Vision*, 1999.

[23] D. N. Wood, A. Finkelstein, J. F. Hughes, C. E. Thayer, and D. H. Salesin. Multiperspective panoramas for cel animation. *Proc. of ACM SIGGRAPH*, pages 243–250, 1997.