Eye Design in the Plenoptic Space of Light Rays

Jan Neumann, Cornelia Fermüller and Yiannis Aloimonos
Computer Vision Laboratory
University of Maryland
College Park, MD 20742-3275, USA
{jneumann,fer,yiannis}@cfar.umd.edu

Abstract

Natural eye designs are optimized with regard to the tasks the eye-carrying organism has to perform for survival. This optimization has been performed by the process of natural evolution over many millions of years. Every eye captures a subset of the space of light rays. The information contained in this subset and the accuracy to which the eye can extract the necessary information determines an upper limit on how well an organism can perform a given task. In this work we propose a new methodology for camera design. By interpreting eyes as sample patterns in light ray space we can phrase the problem of eye design in a signal processing framework. This allows us to develop mathematical criteria for optimal eye design, which in turn enables us to build the best eye for a given task without the trial and error phase of natural evolution. The principle is evaluated on the task of 3D ego-motion estimation.

1. Introduction

Cameras nowadays become smaller and more affordable by the day, thus soon it will be in anyones reach to use assemblies of cameras for the tasks they try to solve. Since these assemblies can be reconfigured easily, the design of new eyes will be possible with little effort. But what is the right design for a given task?

One scientific discipline that has studied this problem is sensory ecology [9]. It studies the relationship between the behavior of an organism, the information it extracts to control its behavior and the environment where it exists.

It has been estimated that eyes have evolved no fewer than forty times, independently, in diverse parts of the animal kingdom. One things that these eye designs and the images that they capture have in common, is that they are highly adapted to the tasks the organism has to perform to survive in its environment. In these organisms all the components of the imaging pipeline from image capture to neuronal encoding and higher level feedback circuits are optimized to increase the fitness of the organism. Examples are the neuronal circuits that encode the captured image in a way that maximizes the amount of information that can be processed with regard to the range of intensities in the environment, or the design differences between the compound eyes of diurnal or nocturnal insects that reflect the tradeoff between light gathering power and visual acuity.

To mimic nature and design a task specific eye, we need to answer the following two questions:

1. What is the relevant visual information that we need to extract to solve our task and how is this information encoded in the visual data that an eye can capture?

2. What is the camera design and image representation that optimally facilitates the extraction of the relevant information?

To answer the first question, we first have to think about what we mean by visual information. When we think about vision, we usually think of interpreting the images taken by (two) eyes such as our own - that is, perspective images acquired by camera-type eyes based on the pinhole principle. These images enable an easy interpretation of the visual information by a human observer. Nowadays though, most processing of visual information is done by machines, thus there is no need to confine oneself to the usual perspective images. Instead we propose to study how the relevant information is encoded in the plenoptic video geometry, that is the geometry of the time-varying space of light rays, to utilize all possible visual information.

To answer the second question we have to determine how well a given eye can capture the necessary information. We can interpret this as an approximation problem where we need to assess how well the relevant subset of the space of light rays can be reconstructed based on the samples captured by the eye, our knowledge of the transfer function of the optical apparatus, and our choice of function space to represent the image. By modeling eyes as spatio-temporal sampling patterns in the space of light rays we can use well developed tools from signal processing and approximation

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

theory to evaluate the suitability of a given eye design for the proposed task and determine the optimal design.

This design approach can result in all-purpose eyes like ours that need to provide input for a variety of different tasks, or it could result in a highly specialized sensor similar to the eyes of stomapods that use up to 16 types of receptors to capture a multi-spectral image of their environment.

Natural evolution suggests a holistic approach to solving visual tasks where hardware and software should be optimized in unison based on the statistics of environment in which the sensors are supposed to operate. The statistics of the space of light rays are hard to capture using conventional imaging devices, but since its brightness structure is very regular we can approximate its statistics from the statistics of natural perspective image sequences which have been studied extensively. This statistical modeling of an "average" environment allows us to design fitness criteria based on mathematical theories that make it possible to short-cut the process of evolution by directly determining the optimal design of a sensor.

1.1. Prior Work

The space of light rays is the most complete visual representation of a scene. It was first studied in the context of photometry and integral photography at the beginning of the 20th century (for an overview see [17]). A mathematical description of the space of light rays is given by the plenoptic function as described by [1]. For each position in space it records the intensity of a light ray for every direction, time, wave length, and polarization, thus providing a complete description of all visual information.

The study of non-perspective subsets of the plenoptic function for vision applications has intensified recently. This is inspired partly by the interest the computer graphics and computer vision community took in using non-perspective subsets of the plenoptic function to represent visual information to be used for image-based rendering such as light fields [13] and lumigraphs [11], or to recover the observed scene from video sequences [5].

This exploration led to studies of new non-perspective imaging geometries (e.g., [12, 18, 19]) and new camera designs that show promise to simplify a large number of vision tasks. Examples are the work on catadioptric sensors [14] for panoramic vision, combinations of filters and lenses for high dynamic range imaging [3], and the design of plenoptic cameras for depth [2, 10] and motion estimation [15]. Despite these advances in the area of eye design, a general framework that relates the design of an imaging sensor to its usefulness for a given task is still missing.

1.2. Outline of Paper

In this work, we attempt to fill this void and propose a new framework for the design of eyes. It consists of the following three steps:

1. Study the structure of the time-varying plenoptic function to determine how the relevant information is encoded.

2. Phrase the eye design problem in a function approximation framework and determine a fitness function that describes how accurately the union of eye design and image processing operators can extract the relevant information from the space of light rays.

3. Compute the optimal eye design by evaluating this fitness function using natural scene statistics and validate the design in the real world.

In the following section, we describe how an eye can be understood as a sampling operator in the space of light rays. Finally, we use a recent result from approximation theory to determine how accurately visual information can be extracted by a given eye design and image representation, and evaluate the approximation error based on natural image statistics. We then demonstrate the proposed framework for eye design by determining the optimal eye for a robot navigating based on visual information.

2. Cameras as Sampling Operators in the Space of Light Rays

In abstract terms, a camera is a mechanism that forms images by focusing light onto a light sensitive surface (retina, film, CCD array, etc.). Different camera designs can be obtained by varying the camera geometry (the geometry of the surface and the geometric distribution of the photoreceptors), and camera optics (the way light is collected and projected onto the surface, e.g. single or multiple lenses, or tubes as in compound eyes, and the optical properties of the photoreceptors). We will use the term polydioptric camera to denote a generalized camera that captures a multi-perspective subset of the space of light rays. We distinguish between the terms *plenoptic*, that denotes the *ideal* concept of all visual information that can possible be captured, and *polydioptric* that denotes the discrete set of visual measurements made by a *physical* sensor to emphasize the notion that all visual information can only be captured with finite precision by any visual sensor.[1]

[1]The word *plenoptic* is derived from the words *complete* and *view*, while *polydioptric* can be loosely translated as "something that is assisting vision by refracting and focusing light in many ways".

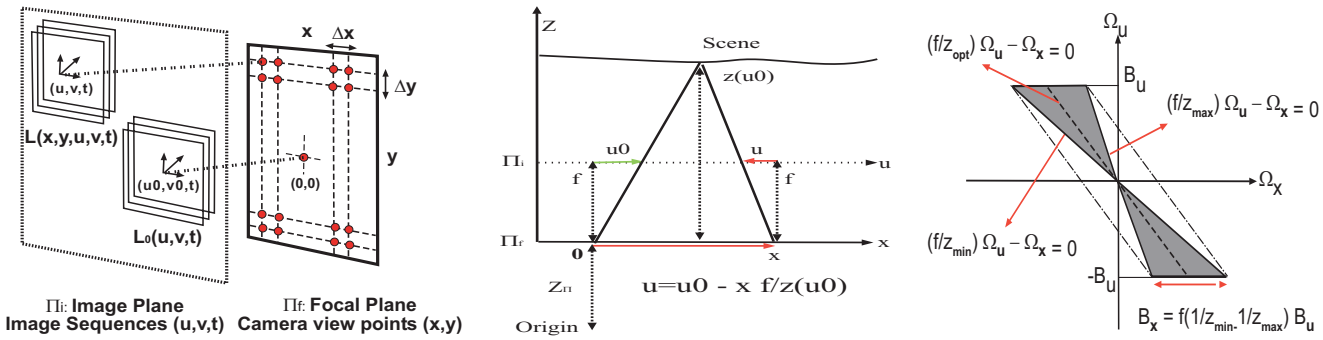

Figure 1. (a) Light field parameterization (b) Light ray correspondence (here shown only for the light field slice spanned by axes x and u). (c) Fourier spectrum of the (x,u) light field slice (epipolar plane) where the depth of scene varies between Z_{min} and Z_{max}, and the perspective image of the scene is band-limited with bounds B_u.

2.1. Parameterization of the Space of Light Rays

At each location $\boldsymbol{x}$ in free space, 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 $\mathcal{L}(\boldsymbol{x};\boldsymbol{r};t)$; $\mathcal{L} : \mathbb{R}^3 \times \mathbb{S}^2 \times \mathbb{R}_+ \rightarrow \Gamma$. Γ denotes here the spectral energy, and equals $\mathbb{R}$ for monochromatic light, $\mathbb{R}^n$ for arbitrary discrete spectra, or could be a function space for a continuous spectrum. $\mathbb{S}^2$ is the unit sphere of directions in $\mathbb{R}^3$.

In a transparent medium such as air the color and intensity of light does not change, thus we can assume that in free space the radiance along the view direction $\boldsymbol{r}$ is constant which implies $\nabla_{\boldsymbol{x}}\mathcal{L}^{\mathrm{T}}\boldsymbol{r} = \nabla_{\boldsymbol{r}}\mathcal{L}^{\mathrm{T}}\boldsymbol{r} = 0$ where $\nabla_{\boldsymbol{x}}\mathcal{L}$ and $\nabla_{\boldsymbol{r}}\mathcal{L}$ are the partial derivatives of $\mathcal{L}$ with respect to position $\boldsymbol{x}$ and direction $\boldsymbol{r}$. 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.

Due to the difficulties involved when using signal processing operators in a mixed spherical-Cartesian coordinate system, we will choose the two-plane parameterization that was used by [11, 13] to represent the space of light rays. All the lines passing through some space of interest can be parameterized by surrounding this space with two nested cubes and then recording the intersection of the light rays 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 we choose both planes to be perpendicular to the z-axis and separated by a distance of f. As seen in Fig. 1a, We denote one plane as *focal plane* Π_f indexed by coordinates (x, y) and the other plane as *image plane* Π_i indexed by (u, v), where (u, v) is defined in a local coordinate system with respect to (x, y). Both (x, y) and (u, v) are aligned with the (X, Y)-axes of the world coordinates and Π_f is at a distance of Z_Π from the origin of the world coordinate system.

This enables us to parameterize the light rays that pass through both planes at any time t using the tuples (x, y, u, v, t) and to record their intensity in the time-varying light field $L(x, y, u, v, t)$.

2.2. Plenoptic Image Formation

A polydioptric camera can be implemented in many ways. The simplest design is an array of ordinary cameras very close to each other (e.g., [23]) or one could use specialized optics or lens systems such as described in [2, 10].

Whatever design one uses, it is not possible to capture the plenoptic function with arbitrary precision. To be able to account for this uncertainty in algorithms at later processing stages, it is important to have an accurate estimate of the approximation error.

Shannon's sampling theorem tells us that a band-limited signal can be recovered exactly when it is sampled at or above the Nyquist rate. On the basis of this [6] examined which rays of a densely captured light field need to be retained to reconstruct the continuous light field without aliasing.

Since natural signals are in general not band-limited in a strict sense, we need to apply a more general sampling theory to our problem. All natural signals have finite energy, thus we can represent the light field as an element of $L_2(\mathbb{R}^5)$, the space of measurable, square integrable functions defined on $\mathbb{R}^5$, and phrase the light field reconstruction problem as a function approximations problem in $L_2(\mathbb{R}^5)$ using recent results in approximation theory [4].

We will use the two-plane parameterization and assume that the imaging elements of the camera sample the light field on a regular lattice in the 5-D space of light rays which corresponds to a choice of camera spacing, image

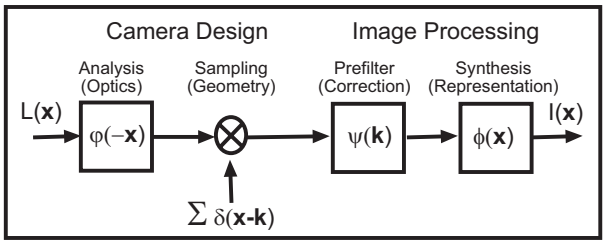

Figure 2. Imaging pipeline expressed in a function approximation framework

resolution, and frame rate. An example setup would be a set of cameras with their optical axes perpendicular to a plane containing the focal points (see Fig. 1a). Then we can describe the 5D periodic lattice $\mathcal{A}$ using 5 vectors $[a_1, a_2, \ldots, a_5]$ which form a lattice matrix A such that $\mathcal{A} = \{A\boldsymbol{k} | \boldsymbol{k} \in \mathbb{Z}^5\}$. A unique tiling of the space of light rays can be achieved by associating with each lattice site a Voronoi cell, which contains all points that are closer to the given lattice site then to any other.

Our model of the image formation pipeline of a polydioptric camera is based on the mathematical framework described in [22] and is summarized in the diagram in Fig. 2.

The camera output is modeled as the inner product of the light field with different translates of an analysis function φ

$$c_\varphi(\boldsymbol{k}) = \int L(\boldsymbol{x})\varphi(\boldsymbol{x} - A\boldsymbol{k})d\boldsymbol{x}; c_\varphi(\boldsymbol{k}) \in l_2(\mathbb{Z}^5) \quad (1)$$

which models the effects of the Pixel Response Function (PRF) such as scattering, blurring, diffraction, flux integration across the pixel's receptive field, shutter time, and other signal degradations. The function $\varphi : \mathbb{R}^5 \rightarrow \mathbb{R}$ is sampled according to the lattice pattern which results in Eq. (1).

The continuous reconstructed light field $I(\boldsymbol{x})$ can be expressed as a linear combination of synthesis functions $\phi : \mathbb{R}^5 \rightarrow \mathbb{R}$ centered on the lattice points, that is

$$I(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in \mathbb{Z}^5} c_\phi(\boldsymbol{k})\phi(\boldsymbol{x} - A\boldsymbol{k});\ c_\phi(\boldsymbol{k}) \in l_2(\mathbb{Z}^5) \quad (2)$$

where l_2 is the space of square-summable sequences $(\sum_{k \in \mathbb{Z}^5} |c(\boldsymbol{k})|^2 < \infty)$.

The coefficients c_ϕ are determined from the camera output c_φ using a linear convolution filter ψ which acts as a correction filter to optimize the projection of the light field signal into the space of images:

$$c_\phi(\boldsymbol{k}) = \sum_i \psi(i)c_\varphi(\boldsymbol{k} - \boldsymbol{i}). \quad (3)$$

How should we choose the filter ψ? In general, since we do not have access to the original signal L, the best we can do is to perpendicularly project the sampling space spanned by the set of analysis functions $\{\varphi(\boldsymbol{x} - \cdot)\}$ on to the image space spanned by the set of synthesis functions $\{\phi(\boldsymbol{x} - \cdot)\}$. This ensures that the image $I(\boldsymbol{x})$ passes through the imaging pipeline unchanged, thus to the camera it "looks" the same as the true light field.

We can also choose the prefilter ψ such that the reconstructed signal interpolates the input signal at the lattice sites, that is $I(A\boldsymbol{k}) = L(A\boldsymbol{k})$. The coefficients of the prefilter can then be determined from the interpolation condition using filter design techniques.

In any case, to evaluate how the error of the light field reconstruction depends on the camera sampling geometry, we compute the average L_2-error $\epsilon^2 = \|L - I\|_{L^2}$ using an integral of the Fourier spectrum of $\hat{L}(\boldsymbol{\Omega})$ and an error kernel $E(\boldsymbol{\Omega})$ [4]:

$$\epsilon^2 \approx \mu^2(A) = \frac{1}{(2\pi)^5} \int_{\mathbb{R}^5} |\hat{L}(\boldsymbol{\Omega})|^2 E(A^{-T}\boldsymbol{\Omega})d\boldsymbol{\Omega} \quad (4)$$

Here A is the sampling lattice matrix, and the error kernel $E(\boldsymbol{\Omega})$ is defined as

$$E(\boldsymbol{\Omega}) = 1 - \frac{|\hat{\phi}(\boldsymbol{\Omega})|^2}{\hat{a}_\phi(\boldsymbol{\Omega})} + \hat{a}_\phi(\boldsymbol{\Omega}) \left| \hat{\tilde{\varphi}}(\boldsymbol{\Omega}) - \frac{\hat{\phi}(\boldsymbol{\Omega})}{a_\phi(\boldsymbol{\Omega})} \right|^2. \quad (5)$$

Here as before ϕ is the synthesis function, $\tilde{\varphi}$ is the the combination of image transfer function φ and prefilter ψ, and a_ϕ is the sampled auto-correlation function of ϕ.

If the light field L is a band-limited then we have the equality $\epsilon = \mu$ for all phase shifts of the signal with respect to the camera, otherwise μ is equal to the average error over all possible phase shifts $L(A\boldsymbol{k} + \boldsymbol{\Delta})$ where $\boldsymbol{\Delta} \in \{A\boldsymbol{x} | \boldsymbol{x} \in [0, 1)^5\}$. This is exactly the expression we are looking for because the relative position of the world with respect to the camera should not matter. Thus, Eq. (4) gives us a means to assess how accurately a given camera design is able to reconstruct the space of light rays in an environment, and thus how accurate we are able to estimate the quantities of interest that we need to compute to solve our task.

2.3. Evaluation of Approximation Error based Natural Image Statistics

To evaluate Eq. 4 we have to choose a synthesis function ϕ, determine the image transfer function φ and compute the appropriate prefilter ψ. In addition, we need to have an idea about the power spectrum of the light field $|\hat{L}(\boldsymbol{\Omega})|^2$. The power spectrum of course depends on the scene in which the sensor operates. If the sensor will only operate in a very constrained environment, then we could capture a number of light fields of this environment, and compute the power spectrum based on this data. This becomes quickly infeasible, if we want to design a visual sensor that performs well in many different environments. Fortunately, to simulate

an "average scene" we can utilize recent studies about the statistics of natural images which we can use (see [20] for a review). The most relevant observation for our purpose is that the power spectrum of a natural image falls approximately inversely proportional to the square of the spatial frequency. Dong and Atick [8] demonstrate how a similar scaling law can be derived from first principles for spatio-temporal sequences. We can use their formalism to find an expression for the power spectrum of a light field.

The power spectrum $P_L(\mathbf{\Omega}) = |\hat{L}(\mathbf{\Omega})|^2$ depends on the spatial frequencies of the textures in the scene, the orientations of the scene surfaces, as well as the depth and velocity distribution of the objects in the scene. For now we will disregard the effect of occlusions and assume that the power spectrum of a perspective image of the scene is rotationally symmetric, that means there is no special direction. For natural scenes this should be roughly satisfied, although it has been observed that horizontal and vertical orientations are often more predominant especially in man-made environments. To simplify the exposition, we will show how only how to determine the power spectrum of a light field subspace formed by the axes u-x-t (time-varying epi-polar plane). The extension to the full 5D light field is straight forward by replacing scalar variables with the corresponding vector variables. For the power spectrum of the static perspective image row we can assume that it follows a power law $P_u(\Omega_u) = |\hat{L}_0(\Omega_u)|^2 = \frac{K}{\|\Omega_u\|^m}$ where $m \approx 2.3$ and K is a normalization constant [8].

We know from [6] that if we disregard occlusions the energy of the Fourier transform of an epipolar plane (x-u-plane) observing an object at a constant depth z is concentrated along the line (we assume unit focal length) $\Omega_x = \Omega_u/z$. Thus the power spectrum of a static epipolar plane is given by $P_u(\Omega_u)\delta(\Omega_x - \Omega_u/z)$. If the depth is varying, then the power spectrum will spread out to a wedge-shaped region bounded by the minimal and maximal depth (see Fig 1c). For a given image region that moves in the image plane with velocity (optical flow) $\dot{x}$, the power spectrum of a perspective spatio-temporal image plane is concentrated along the line $\Omega_t = \Omega_u\dot{x}$. By combining these two constraints, we can write the power spectrum of the time-varying epipolar plane as follows:

$$P_L(\Omega_x, \Omega_u, \Omega_t, z, \dot{\boldsymbol{x}}) \tag{6}$$
$$= P_u(\Omega_u)\delta(\Omega_x - \Omega_u/z)\delta(\Omega_t - \Omega_u\dot{x}) \tag{7}$$

Given a probability distribution for the velocities $D_{\dot{x}}(\dot{x})$ and depths $D_z(z)$, and using the fact that the term under the integral is only nonzero (due to the delta functions) for

$$z = \Omega_u/\Omega_x \text{ and } \dot{x} = \Omega_t/\Omega_u, \tag{8}$$

we can express an "average" light field power spectrum by integrating over these distributions as:

$$|\hat{L}(\Omega_x, \Omega_u, \Omega_t)|^2 = P_u(\Omega_u)\cdot$$
$$\int_0^\infty \int_0^\infty \delta(\Omega_x - \frac{\Omega_u}{z})\delta(\Omega_t - \Omega_u\dot{x})D_z(z)D_{\dot{x}}(\dot{x})dzd\dot{x}$$
$$= \frac{K}{\|\Omega_u\|^m} D_{\dot{x}}(\frac{\Omega_t}{\Omega_u})D_z(\frac{\Omega_u}{\Omega_x}). \tag{9}$$

This expression allows us now to evaluate Eq. (4) given a depth and velocity distribution. In conclusion, by combining Eqs. (9) and (4) we can evaluate how accurately a given camera assembly is able to capture the light field of an "average scene". When this error estimate is combined with a sensitivity analysis of the vision algorithm that will operate on the images, then we can define a mathematical fitness function over the space of camera assemblies, and by optimizing over the camera geometry (as defined by the lattice matrix A) and the camera optics (analysis function φ and prefilter ψ), we can determine the optimal camera for the task at hand.

3. Case Study: Designing Eyes for a Robot

In this section we will demonstrate the camera design methodology as described in the previous sections, by designing eyes for a robot that needs to navigate based on visual information. We use 3D motion estimation as an example task because it is one of the fundamental problems in vision and a fundamental component of many algorithms in navigation, virtual reality, tele-immersion, and graphics.

To determine the optimal eye for the robot, we first have to analyze the brightness structure of the plenoptic function to determine what plenoptic subspace the robot eye should capture.

3.1. How is 3D motion information encoded in the space of light rays?

In a static world the brightness structure of the space of light rays is time-invariant, thus if a camera moves rigidly and captures two overlapping sets of light rays at two different time instants, then a subset of these rays should match exactly and would allow us to recover the rigid motion from the light ray correspondences. If we choose the camera coordinate system as our fiducial coordinate system, we can describe this motion by an opposite rigid coordinate transformation of the ambient space of light rays in the camera coordinate system. This rigid transformation, parameterized by the rotation matrix $R(t)$ and a translation vector $\boldsymbol{q}(t)$, results in the following *exact* equality which is called the *discrete plenoptic motion constraint* [16]:

$$\mathcal{L}(R(t)\boldsymbol{x} + \boldsymbol{q}(t); R(t)\boldsymbol{r}; t) = \mathcal{L}(\boldsymbol{x}; \boldsymbol{r}; 0). \tag{10}$$

Subspace Dimension	Subspace Axes	Example Motion Constraint Equation (zero columns imply that no constraint exists for that parameter)	Parameters to estimate	Need to estimate depth?
2D:	xt,yt ut,vt	$-L_t = L_u \begin{pmatrix} \frac{f}{z} & 0 & -\frac{u}{z} & 0 & \frac{u^2}{f}+f & 0 \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{q}} \\ \boldsymbol{\omega} \end{pmatrix}$	2+N	yes
3D:	xyt,xvt yut,uvt	$-L_t = \begin{pmatrix} L_u \\ L_v \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \frac{f}{z} & 0 & -\frac{u}{z} & -\frac{uv}{f} & \frac{u^2}{f}+f & -v \\ 0 & \frac{f}{z} & -\frac{v}{z} & -(\frac{v^2}{f}+f) & \frac{uv}{f} & u \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{q}} \\ \boldsymbol{\omega} \end{pmatrix}$	5+N	yes
3D:	xut,yvt	$-L_t = \begin{pmatrix} L_x \\ L_u \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 0 & -\frac{u}{f} & 0 & \frac{ux}{f}+Z_\Pi & 0 \\ 0 & 0 & 0 & 0 & \frac{u^2}{f}+f & 0 \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{q}} \\ \boldsymbol{\omega} \end{pmatrix}$	3	no
4D:	xyut,xyvt xuvt,yuvt	$-L_t = \begin{pmatrix} L_x \\ L_u \\ L_v \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 0 & -\frac{u}{f} & 0 & \frac{ux}{f}+Z_\Pi & 0 \\ 0 & 0 & 0 & -\frac{uv}{f} & \frac{u^2}{f}+f & -v \\ 0 & \frac{f}{z} & -\frac{v}{z} & -(\frac{v^2}{f}+f+\frac{f}{z}Z_\Pi) & \frac{vu}{f}+\frac{vx}{z} & u+\frac{f}{z}x \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{q}} \\ \boldsymbol{\omega} \end{pmatrix}$	6 + N	yes
5D:	xyuvt	$-L_t = \begin{pmatrix} L_x \\ L_y \\ L_u \\ L_v \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 0 & -\frac{u}{f} & -\frac{uy}{f} & \frac{ux}{f}+Z_\Pi & -y \\ 0 & 1 & -\frac{v}{f} & -(\frac{vy}{f}+Z_\Pi) & \frac{vx}{f} & x \\ 0 & 0 & 0 & -\frac{uv}{f} & \frac{u^2}{f}+f & -v \\ 0 & 0 & 0 & -(\frac{v^2}{f}+f) & \frac{vu}{f} & u \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{q}} \\ \boldsymbol{\omega} \end{pmatrix}$	6	no

Table 1. Information content of the different plenoptic subspaces with regard to the 3D motion estimation problem

It expresses the fact that the rigid motion maps the time-invariant space of light rays upon itself.

As was shown in [16], if the motion of the camera is small, then we can express the differential changes in a spatio-temporal light field in terms of the light field coordinates (x, y, u, v, t) and light field derivatives $L_x = \partial L/\partial x, \ldots$. ($[\cdot;\cdot]$ denotes the vertical stacking of vectors):

$$-L_t = [L_x, L_y, L_u, L_v][M_t, M_\omega][\dot{\boldsymbol{q}}; \boldsymbol{\omega}] \tag{11}$$

where

$$M_t = \begin{pmatrix} 1 & 0 & -\frac{u}{f} \\ 0 & 1 & -\frac{v}{f} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; M_\omega = \begin{pmatrix} -\frac{uy}{f} & \frac{ux}{f}+Z_\Pi & -y \\ -(\frac{vy}{f}+Z_\Pi) & \frac{vx}{f} & x \\ -\frac{uv}{f} & \frac{u^2}{f}+f & -v \\ -(\frac{v^2}{f}+f) & \frac{vu}{f} & u \end{pmatrix}.$$

This linear constraint equation allows us to compute the motion of the camera directly from the captured image information by solving a simple over-determined set of linear equations in the 6 rigid motion parameters.

In both the discrete and the differential case we see that if a sensor is able to capture a continuous non-degenerate subset of the plenoptic function, then the problem of estimating the rigid motion of this sensor is *independent of the scene* and the only free parameters are the six degrees of freedom of the rigid motion. This global parameterization leads to a highly constrained estimation problem.

One question comes to mind. Do we truly need to capture the full plenoptic function to utilize this constraint? Unfortunately, the answer is yes. If we only have access to a lower-dimensional subspace of the light field, we can only form reduced motion constraint equations from Eq. (11) by omitting the dimensions along which we cannot make measurements or utilizing the triangulation relation ship between the flow in the focal and image planes ($[\dot{u}; \dot{v}] \cdot z(x, y, u, v) = f[\dot{x}; \dot{v}]$ as illustrated in Fig. 1b). We collected the motion constraint equations for all the plenoptic subspaces in Table 1, and we see that the camera that makes the motion estimation problem the easiest is the one that samples the whole plenoptic function (or a multi-perspective 3D slice of it for the case of planar motion) because the motion estimation problem is reduced to a low-dimensional image registration problem as said before.

Another important criteria is the range of directions (field of view) of the sensor. A small field of view makes the motion estimation ill-posed (see [7] for a study on this subject), thus for accurate and robust motion estimation the sensor needs to have a wide field of view. Combining the two criteria, the field of view and the subset of the space of light rays that a sensor captures, we can rank different eye design in a hierarchy as shown in Fig. 3 which expresses a qualitative measure of how hard the task of motion estimation is to solve for a given sensor design [16].

3.2. What Camera Design Can Best Extract the Motion Information?

The qualitative hierarchy of camera designs in Fig. 3a can be quantitatively analyzed by applying the framework presented in section 2. To simplify the exposition we assume that the robot is only able to move on a planar, flat surface, thus the locomotion of the robot is limited to a horizontal planar motion, and that the camera designs under study are restricted sets of horizontally aligned pinhole line cameras. As summarized in Table 1 in the previous section, we can see that we can extract the 3 planar motion parameters directly from the image data if we are able to capture

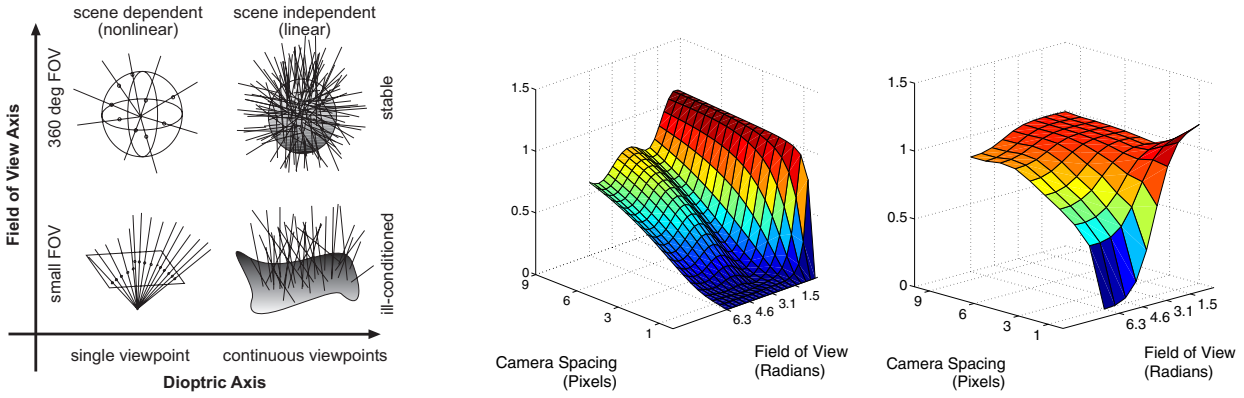

Figure 3. Hierarchy of Cameras for 3D Motion Estimation. (a) Qualitative Hierarchy: The different camera models are classified according to the field of view (FOV) and the number and proximity of the different viewpoints that are captured (Dioptric Axis). The camera models are clockwise from the lower left: small FOV pinhole camera, spherical pinhole camera, spherical polydioptric camera , and small FOV polydioptric camera. (b) Metric on Space of Cameras for the "theoretically average" environment: Relative error planar rotation estimate in dependence on camera spacing and field of view of cameras. The plot was generated by evaluating Eqs. (4) **and** (9)**. (c) Metric on Space of Cameras for an empirical environment: The plot was generated by averaging the estimation errors over many line-image sequences generated from an epipolar volume.**

the full x-u-t subset of the light field surrounding the robot. If we only have a single line camera, then we can only compute the heading direction and explicitly need to estimate the depth of the scene.

Given such a hardware setup, how accurately can we approximate the x-u-t subspace of the light field and especially its derivatives? It was shown in [4] that cubic B-splines offer the best compromise between order of approximation and and minimal support, thus we will choose them as our synthesis basis function ϕ. For the camera itself we will assume that the camera is an ideal acquisition device, thus we choose ψ as the B-splines prefilter. For this case [4] give an explicit formula of the error kernel for the case of cubic B-Splines. To evaluate how accurate we can compute the image derivatives, we need to replace the light field L and the synthesis function ϕ in Eq. (4) by their respective derivatives. Then we choose a distribution of depths and velocities in the scene, which leaves the sampling lattice A of the camera assembly as the variable to optimize over. Since there is a linear relationship between differential light field derivatives and the rigid motion parameters, the sensitivity analysis of the motion estimation can be evaluated by utilizing results from the perturbation theory of the linear least-squares problem [21] where the distribution of the errors depends on the approximation error computed using Eq. (4). In Fig. 3b we plot the relative error in the estimate of the rotation angle of the planar camera motion for varying field of view and distance between the camera centers.

Another approach to simulating and evaluating camera designs without actually building them is based on resampling previously recorded subsets of the space of light rays. We captured a number of epipolar volumes [5] of different scenes which varied in depth and texture complexity. Given such a continuous subset of the plenoptic function we are able to generate new line-camera image sequences by resampling this set of light rays. These generated image sequences are essentially identical to the image sequences that a true physical line camera would capture as long as the camera motion is chosen such that all pixels of the camera can be interpolated from the voxels of the epipolar plane volume. By varying the spatio-temporal sampling pattern we can simulate a wide range of camera motions as well as camera designs. We generated a large number of image sequences for various camera motions and distances between the camera centers. For each frame of a sequence we formed the plenoptic motion constraint equations (consisting in this case of the rows of Eq. (11) corresponding to a planar motion) and solved for the planar motion parameters using the plenoptic derivatives. As an example result, we show in Fig. 3c how the accuracy of the rotation estimate improves when the field of view increases, and how the accuracy decreases when the spacing between the cameras increases. As expected, the empirical metric on the space of cameras is qualitatively similar to the theoretically derived average metric in Fig 3b, although the exact shape of the error surface is not very similar. Whether the theoretical

or the empirical approach to defining a metric on the space of cameras will be more useful for designing new sensors needs to be further examined and is subject of current work.

4. Summary and Conclusions

In this paper we presented a new methodology for the design of eyes which interprets camera assemblies as sampling operators in the space of light rays. This opens up the treasure chest of signal processing tools to aid in the design of new cameras, and we showed how the problem of camera design can be rephrased as a filter optimization problem. This optimization problem can then be solved by utilizing domain knowledge such a natural image statistics and depth and velocity distributions in the scene, or by simulating the behavior of the sensor using real plenoptic data such es epipolar volumes or light fields. The methodology was demonstrated by assessing the influence of two camera design parameters on the accuracy of ego motion estimation. We believe that rephrasing the problem of camera design in terms of a function approximation framework has great potential especially with advent of optical nano-technology around the corner which will offer new opportunities to design revolutionary different cameras that sample the space of light rays in ways unimaginable to us today.

Acknowledgements

The support through the National Science Foundation Award 0086075 is gratefully acknowledged.

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] Manoj Aggarwal and Narendra Ahuja. High dynamic range panoramic imaging. In *Proc. Int. Conf. Computer Vision*, pages 2–9, 2001.

[4] T. Blu and M. Unser. Quantitative Fourier analysis of approximation techniques: Part I—Interpolators and projectors. *IEEE Transactions on Signal Processing*, 47(10):2783–2795, October 1999.

[5] 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.

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

[7] K. Daniilidis and M. Spetsakis. Understanding noise sensitivity in structure from motion. In *Visual Navigation: From Biological Systems to Unmanned Ground Vehicles*, chapter 4, pages 61–88. Lawrence Erlbaum Associates, Hillsdale, NJ, 1997.

[8] D.W. Dong and J.J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345–358, 1995.

[9] D.B. Dusenbery. *Sensory Ecology*. W.H. Freeman and Company, New York, 1992.

[10] H. Farid and E. Simoncelli. Range estimation by optical differentiation. *Journal of the Optical Society of America*, 15(7):1777–1786, 1998.

[11] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. In *Proceedings of ACM SIGGRAPH 96*, Computer Graphics (Annual Conference Series), pages 43–54, New York, 1996. ACM, ACM Press.

[12] 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.

[13] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of ACM SIGGRAPH 96*, Computer Graphics (Annual Conference Series), pages 161–170, New York, 1996. ACM, ACM Press.

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

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

[16] J. Neumann, C. Fermüller, and Y. Aloimonos. Polydioptric camera design and 3d motion estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 294–301, 2003.

[17] T. Okoshi. *Three-dimensional Imaging Techniques*. Academic Press, 1976.

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

[19] S. Seitz. The space of all stereo images. In *Proc. International Conference on Computer Vision*, pages 307–314, 2001.

[20] E.P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, May 2001.

[21] G.W. Stewart. Stochastic perturbation theory. *SIAM Review*, 32:576–610, 1990.

[22] M. Unser and A. Aldroubi. A general sampling theory for nonideal acquisition devices. *IEEE Transactions on Signal Processing*, 42(11):2915–2925, November 1994.

[23] Bennett Wilburn, Michael Smulski, Hsiao-Heng Kelin Lee, and Mark Horowitz. The light field video camera. In *Proceedings of Media Processors*. SPIE Electronic Imaging, 2002.