Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2008-021 April 16, 2008

Understanding camera trade-offs through a Bayesian analysis of light field projections

Anat Levin, William T. Freeman, and Fredo Durand

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Understanding camera trade-offs through a Bayesian analysis of light field projections

Anat Levin William T. Freeman Frédo Durand

Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory

Abstract. Computer vision has traditionally focused on extracting structure, such as depth, from images acquired using thin-lens or pinhole optics. The development of computational imaging is broadening this scope; a variety of unconventional cameras do not directly capture a traditional image anymore, but instead require the joint reconstruction of structure and image information. For example, recent coded aperture designs have been optimized to facilitate the joint reconstruction of depth and intensity. The breadth of imaging designs requires new tools to understand the tradeoffs implied by different strategies.
This paper introduces a unified framework for analyzing computational imaging approaches. Each sensor element is modeled as an inner product over the 4D light field. The imaging task is then posed as Bayesian inference: given the observed noisy light field projections and a new prior on light field signals, estimate the original light field. Under common imaging conditions, we compare the performance of various camera designs using 2D light field simulations. This framework allows us to better understand the tradeoffs of each camera type and analyze their limitations.

1 Introduction

The flexibility of computational imaging has led to a range of unconventional designs that facilitate structure inference and post-processing. Cameras with coded apertures [1,2,3], plenoptic cameras [4,5,6], phase plates [7,8], stereo [9], multi-view systems [10,11,12], depth from defocus systems [13,14,15,16,17,18,19,20,21,22,23,24,25], radial catadioptric imaging [26], lensless imaging [27], mirror arrays [28,29], or even random cameras [29,30] all record different combinations of the light rays. Reconstruction algorithms based on a combination of signal processing and machine vision then convert the data to viewable images, potentially with richer information such as depth or a full 4D light field. Each of these cameras involves tradeoffs along various dimensions –spatial and depth resolution, depth of focus and noise sensitivity. This paper describes a theoretical framework that will help us to compare computational camera designs and understand their tradeoff in terms of image and structure inference.

Computation is changing imaging in three fundamental ways. First, the information recorded at the sensor may not be the final image, and the need for a decoding algorithm must be taken into account to assess camera quality. Second, the output and intermediate data are not limited to flat 2D images anymore and new designs enable the extraction of 4D light fields and depth information. Finally, new *priors* or statistical models can capture regularities of natural scenes to complement the sensor measurements and amplify

the power of decoding algorithms. The traditional evaluation tools based on image PSF and frequency responses [31,32] are not able to fully model these effects. Our goal in this paper is to develop tools for a comparison across different imaging designs, taking into account those three aspects. We want to evaluate the ability to recover a 2D image as well as depth or other information. We want to model the need for a decoding step and the use of natural-scene priors.

Given the variety of designs and types of information, we argue that a powerful common denominator is the notion of light field [10] because it directly encodes light rays- the atomic entities interacting with the camera sensor. Light fields naturally encapsulate some of the more common photography goals such as high spatial image resolution, and are tightly coupled with the targets of mid-level computer vision: surface depth, texture, and illumination information. This means that we need to cast the reconstruction performed in computational imaging as a light field inference problem. In order to benefit from recent advances in computer vision, we also need to extend prior models, traditionally studied for 2D images, to 4D light fields.

In a nutshell, the operation of camera sensors can be modeled as the integration of a set of light rays, with the optics specifying the mapping between rays and sensor elements. Thus, in an abstract way, a camera provides a linear projection of the 4D light field where each coordinate corresponds to the measurement of one pixel. The goal of a decoding process is to infer from such projections as much information as possible about the 4D light field. Since the number of sensor elements is significantly smaller than the dimensionality of the light field signal, prior knowledge on light fields is essential. We analyze the limitations of traditional signal processing assumptions [33,34,35,36,37] and suggest a new prior on light field signals which explicitly accounts for their locally elongated structure. We then define a new metric of camera performance as follows: Given a light field prior, from the data measured by the camera, how well can the light field be reconstructed? The number of sensor elements is of course a critical variable, and the evaluations in this paper are normalized by imposing a fixed budget of N sensor elements to all cameras. This is not a strict requirement of our approach, but it provides a meaningful common basis.

Our evaluation focuses on the information captured by a projection, omitting the confounding effect of camera-specific inference algorithms. We also do not address decoding complexity. For clarity of exposition and computational efficiency we focus on the 2D version of the problem (1D image/2D light field). We use simplified optical models and do not model lens aberrations or diffraction. These effects would still follow a linear projection model and can be accounted for with modifications to the light field projection function.

Using light fields generated by ray tracing, we simulate several existing projections (cameras) under equal conditions, and demonstrate the quality of reconstruction they can provide.

Our framework captures the three major elements of the computational imaging pipeline – optical setup, decoding algorithm, and priors – and enables a comparison on a common baseline. This framework allows us to systematically compare computational camera designs at one of the most basic computer vision task: estimating the light field from sensor responses.

1.1 Related Work

Approaches to lens characterization such as Fourier Optics and MTF [31,32] analyze an optical element in terms of signal bandwidth and the sharpness of the PSF over the depth of field, but do not address depth information. The growing interest in 4D light field rendering has led to research on reconstruction filters and anti-aliasing in 4D [33,34,35,36,37], yet this research relies mostly on classical signal processing assumptions of band limited signals, and do not utilize the rich statistical correlations of light fields. Research on generalized camera families [38,39,40] mostly concentrates on geometric properties and 3D configurations, but with an assumption that approximately one light ray is mapped to each sensor element and thus decoding is not taken into account. In [41] aperture effects were modeled but decoding and information were not yet analyzed.

Reconstructing data from linear projections is a fundamental component in tools such as CT and tomography [42]. Fusing multiple image measurements is also used for super-resolution, and [43] studies inherent uncertainties in this process. In [44], the concept of compressed sensing is used to study the ability to reconstruct a signal from arbitrary random projections, when the signal is sufficiently sparse in some representation. Weiss et al [45] attempt to optimize such projections. While sparsity is a stronger statistical assumption than band limited signals, it still does not capture many structural aspects of light fields.

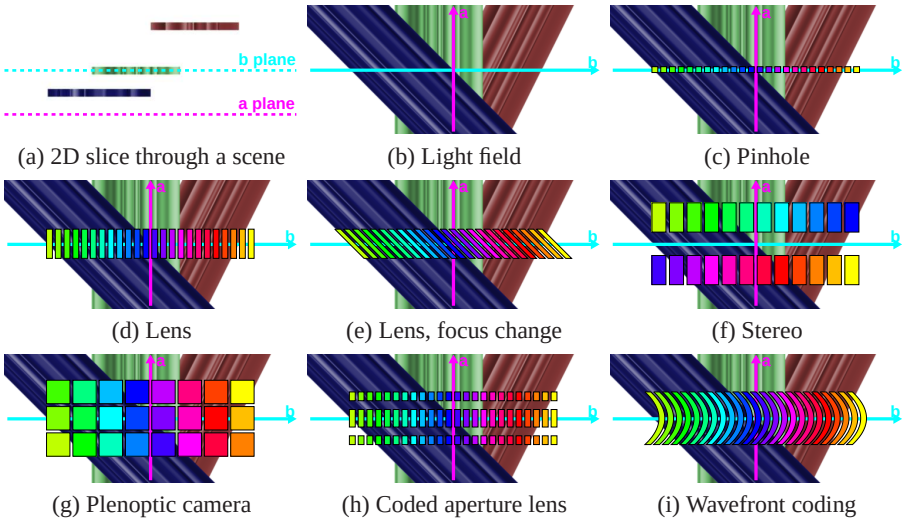
2 Light fields and camera configurations

Light fields are 4D functions that encode the radiance for each light ray leaving a scene. Light fields are usually represented with a two-plane parameterization, where each ray is encoded by its intersections with two parallel planes. Figure 1(a,b) shows a 2D slice through a diffuse scene and the corresponding 2D slice out of the 4D light field. The color at position (a_0, b_0) of the light field in fig. 1(b) is that of the reflected ray in fig. 1(a) which intersects the **a** and **b** lines at points a_0, b_0 respectively. Each row in this light field corresponds to a 1D view when the viewpoint shifts along **a**. One of the most distinctive properties of light fields is the strong elongated lines. For example the green object in fig. 1 is diffuse and the reflected color does not vary along the **a** dimension. Specular objects exhibit some variation along the **a** dimension, but typically much less than along the **b** dimension. The slope of those lines encodes the object's depth, or disparity [33,34].

Each sensor element records the amount of light collected from multiple rays and can be thought of as a linear sum over some set of light rays. For example, in a conventional lens, the value at a pixel is an integral of rays over the lens aperture and the sensor photosite. We review several existing camera configurations and express the rule by which they project light rays to sensor elements. We assume that the camera aperture is positioned on the **a** line parameterizing the light field.

Ideal Pinhole cameras Each sensor element collects light from a single ray, and the camera projection just slices a row in the light field (fig 1(c)). Since only a tiny fraction of light is let in, noise is an issue.

Lenses Lenses can gather more light by focusing all light rays emerging from a point at a given distance D to a single sensor point. In the light field, $1/D$ is the slope



(a) 2D slice through a scene (b) Light field (c) Pinhole

(d) Lens (e) Lens, focus change (f) Stereo

(g) Plenoptic camera (h) Coded aperture lens (i) Wavefront coding

Fig. 1. (a) Flat-world scene with 3 objects. (b) The light field, and (c)-(i) cameras and the light rays integrated by each sensor element (distinguished by color)

of the integration (projection) stripe (fig 1(d,e)). An object is in focus when its slope matches this slope (e.g. the green object in fig 1(d)) [33,34,35,36]. Objects in front or behind the focus distance will be blurred. Larger apertures gather more light but cause more defocus.

Stereo Stereo pairs [9] facilitate depth inference, by recording two views of the scene (fig 1(g), to maintain a constant sensor element budget, the resolution of each image is halved).

Plenoptic cameras To capture multiple viewpoints, plenoptic cameras use a microlens array between the lens and the sensor [4,5]. These microlenses separate the rays according to their direction, thereby recording many samples of the full 4D light field impinging the main lens. If each microlens covers k sensor elements, one achieves k different views of the scene, but the spatial resolution is reduced by a factor of k ($k = 3$ is shown in fig 1(g)).

Coded aperture Recent work [2,3] places a code at the lens aperture, blocking light rays (fig 1(h)). As with conventional lenses, objects deviating from the focus depth are blurred, but according to the aperture code. The code is designed to be highly sensitive to scale variations. Since the blur scale is a function of depth, by searching for the code scale which best explains the local image window, depth can be inferred. Given depth, the blur can also be inverted, increasing the depth of field.

Wavefront coding introduces an optical element with an unconventional shape (phase plate) so that rays from any world point do not converge to a single sensor element [7][1]. This can be thought of as integrating over a curve in light field space (see fig 1(i)), instead of the straight strip integration of lenses. This makes the defocus of different depths almost identical, which enables deconvolution without depth information,

[1] While wavefront coding is usually derived in terms of wave otpics, the resulting system is usually illustrated with ray diagrams.

thereby extending depth of field. To achieve this, a cubic lens shape (or phase plate) is used and the light field integration curve, and the derivative of the cubic surface is parabolic. Since the integration curve is a function of the lens normal, it is parabolic as well (fig 1(i)).

3 Bayesian estimation of light field

3.1 Problem statement

We model an imaging process as an integration of light rays by camera sensors, or in an abstract way, as a linear projection of the light field

$$y = Tx + n \tag{1}$$

where x is the light field, y is the captured image, n is an iid Gaussian noise $n \sim N(0, \eta^2 I)$ and T is the projection matrix, describing how light rays are mapped to sensor elements. Referring to figure 1, T includes one row for each sensor element, and this row has non-zero elements for the light field entries marked by the corresponding color (e.g. a pinhole T matrix has a single non-zero element per row).

The set of realizable T matrices is limited by physical constraints. In particular, the entries of the projection matrix T are all non-negative. To ensure equal conditions for noise issues, we assume that a maximal integration time is allowed, and normalize it so that the maximal value for each entry of T is 1. The total amount of light reaching each sensor element is the sum of the entries in the corresponding T row. It is usually desired to collect more light to increase the signal to noise ratio. For example, a pinhole is noisier because it has a single non-zero entry per row, while a lens has multiple ones.

To simplify notation, most of the following derivation will address a 2D slice in the 4D light field, but the 4D case is similar. While the light field is naturally continuous, for simplicity we use a discrete representation.

Our goal is to understand how well we can recover the light field x from the noisy projection y, and which T matrices, among the list of camera projections described in the previous section, permit better reconstructions. That is, if one is allowed to take N measurements (T can have N rows), which set of projections leads to better light field reconstruction? Our evaluation metric can be adapted to a weight field w which specifies how much we care about reconstructing different parts of the light field. For example, if the goal is an all-focused, high quality image from a single view point (as in wavefront coding), we can assign zero weight to all but one light field row.

The number of measurements taken by most optical systems is significantly smaller than the light field data, or in other words, the projection matrix T contains many fewer rows than columns. This makes the recovery of the light field ill-posed and motivates the use of prior knowledge on the generic structure of light fields. We therefore start by asking how to model a light field prior.

3.2 Classical priors

State of the art light field sampling and reconstruction approaches [33,34,35,36,37] apply signal processing techniques, which are mostly based on band-limited signal assumptions. The principle is that the number of non-zero frequencies in the signal has to be equal to the number of samples. Thus, before samples are taken, one has to apply a

low-pass filter to meet the Nyquist limit. Light field reconstruction is then reduced to a convolution with a proper low-pass filter. When the depth range in the scene is bounded, these strategies can further bound the set of active frequencies within a sheared rectangle instead of a standard square of low frequencies and tune the orientation of the low pass filter. They also provide principled rules for trading spatial and directional samples. However, they focus on pure sampling/reconstruction approaches and do not address inference for a general projection such as the coded aperture.

One way to express the underlying band limited assumptions in a prior terminology is to think of an isotropic Gaussian prior. In the frequency domain, the covariance of such a Gaussian is diagonal, allowing a very narrow variance at the highest frequencies, and a wider one at the lower frequencies. Similar priors can also be expressed in the spatial domain by penalizing the convolution with a set of high pass filters:

$$P(x) \propto exp(-\frac{1}{2\sigma_0}\sum_{k,i}|f_{k,i}x^T|^2) = exp(-\frac{1}{2}x^T\Psi_0^{-1}x) \tag{2}$$

where $f_{k,i}$ denotes the kth high pass filter centered at the ith light field entry. In sec 5, we will show that band limited assumptions and Gaussian priors indeed lead to equivalent sampling conclusions.

An additional option is to think of a more sophisticated high pass penalty and replace the Gaussian prior of eq 2 with a heavy-tailed prior [46].

However, as will be illustrated in section 3.4, such generic priors ignore the very strong elongated structure of light fields, or the fact that the variance along the disparity slope is significantly smaller than the spatial variance.

3.3 Mixture of Gaussians (MOG) Light field prior

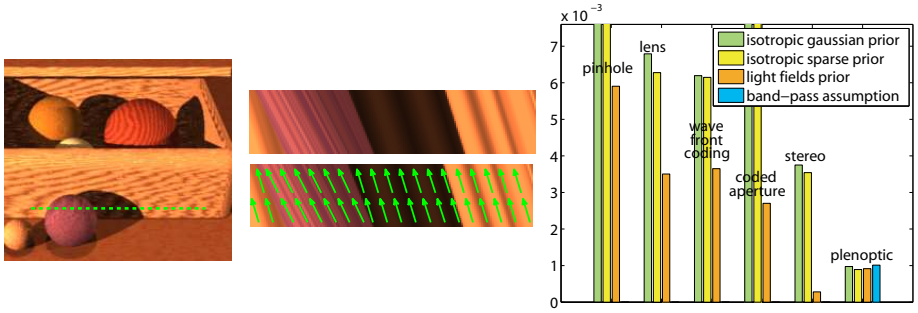
To account for the strong elongated structure of light fields, we propose modeling a light field prior using a mixture of oriented Gaussians, where each Gaussian component corresponds to a depth interpretation of the scene. If the scene depth (and hence light field slope) is known we can define an anisotropic Gaussian prior that accounts for the oriented structure. For this, we define a slope field S that represent the slope (one over the depth of the visible point) at every light field entry (fig. 2(b) illustrates a sparse sample from a slope field). For a given slope field, our prior assumes that the light field is Gaussian, but has a variance in the disparity direction that is significantly smaller than the spatial variance. The covariance Ψ_S corresponding to a slope field S is then:

$$x^T\Psi_S^{-1}x = \sum_i \frac{1}{\sigma_s}|g_{S(i),i}^T x|^2 + \frac{1}{\sigma_0}|g_{0,i}^T x|^2 \tag{3}$$

where $g_{s,i}$ is a derivative filter in orientation s centered at the ith light field entry (specifically $g_{0,i}$ is the derivative in the horizontal/spatial direction), and $\sigma_s << \sigma_0$, especially for specular objects. Conditioning on depth we have $P(x|S) \sim N(0, \Psi_S)$.

We also need a prior $P(S)$ on the quality of a slope field S. Given that depth is usually piecewise smooth, our prior encourages piecewise smooth slope fields (like the depth regularization of conventional stereo algorithms). Note however that S and this prior are expressed in light-field space, not image or object space. The resulting unconditional light field prior is an infinite mixture of Gaussians (MOG) that sums over slope fields

$$P(x) = \int_S P(S)P(x|S) \tag{4}$$



(a) Test image (b) light field and slope field (c) SSD error in reconstruction

Fig. 2. Light field reconstruction.

We note that while each mixture component is a Gaussian which can be evaluated in closed form, marginalizing over the infinite set of slope fields S is intractable, and approximation strategies are described below.

Now that we have modeled the probability of a light field x to be "natural", we turn to the imaging problem: Given a camera T and a noisy projection y we want to find a Bayesian estimate for the light field x. For this, we need to define $P(x|y;T)$, the probability that x is the explanation of the measurement y. Using Bayes' rule:

$$P(x|y;T) = \int_S P(x,S|y;T) = \int_S P(S|y;T)P(x|y,S;T) \tag{5}$$

To express the individual terms in the above equation, we note that y should be equal to Tx up to measurement noise, that is, $P(y|x;T) \propto exp(-\frac{1}{2\eta^2}|Tx-y|^2)$. As a result, for a given slope field S, $P(x|y,S;T) \propto P(x|S)P(y|x;T)$ is also Gaussian with covariance and mean:

$$\Sigma_S^{-1} = \Psi_S^{-1} + \frac{1}{\eta^2}T^TT \qquad \mu_S = \frac{1}{\eta^2}\Sigma_S T^T y \tag{6}$$

Similarly, $P(y|S;T)$ is also a Gaussian distribution measuring how well we can explain y with the slope component S, or, the volume of light fields x which can explain the measurement y, if the slope field was S. This can be computed by marginalizing over light fields x: $P(y|S;T) = \int_x P(x|S)P(y|x;T)$. Finally, $P(S|y;T)$ is obtained with Bayes' rule: $P(S|y;T) = P(S)(y|S;T)/\int_S P(S)(y|S;T)$

To recap, Since we model our light field prior as a mixture of Gaussians conditioned on a slope field, the probability $P(x|y;T)$ that a light field x explains a measurement y is also a mixture of Gaussians (MOG). To evaluate it, we measure how well x can explain y, conditioning on a particular slope field S, and weigh it by the probability $P(S|y)$ that S is actually the slope field of the scene.

Inference Given a camera T and an observation y our goal is to infer a MAP estimate of x, but the integral in eq 5 is intractable. Our strategy is to approximate the MAP estimate for the slope field S, and conditioning on this estimate, solve for the MAP light field.

The slope field inference stage is essentially inferring the unknown scene depth. Our inference generalizes MRF stereo algorithms [9] or the depth regularization of the coded aperture approach [2]. The exact details about slope inference are provided in

the appendix, but as a brief summary, we model slope in local windows as constant or having one single discontinuity, and we then regularize the estimate using a MRF.

Given the estimated slope field S, our light field prior is Gaussian, and thus the MAP estimate for the light field is the mean of the conditional Gaussian μ_S in eq 6. This mean will attempt to minimize the projection error up to noise, and regularize the estimate by attempting to minimize the oriented variance Ψ_S. Note that in traditional stereo formulations the multiple views are used only for depth estimate. In contrast, the formulation of our light field estimate seeks a light field that will satisfy the projection in all views. Thus, if the individual views include aliasing, we can achieve "super resolution".

3.4 Empirical illustration

To illustrate the light field inference, figure 2(a,b) presents an image and a light field slice, involving depth discontinuities. Fig 2(c) presents the numerical SSD estimation errors. Figures 3,4 presents visually the estimated light fields and (sparse samples from) the corresponding slope fields. See supplementary file for more results. Note that slope errors often accompany ringing in the reconstruction. We compare the results of the MOG light field prior with simpler Gaussian priors (extending the conventional band limited signal assumptions [33,34,35,36,37]) and with modern sparse derivative priors [46,44]. For the plenoptic camera case we also explicitly compare with the signal processing reconstruction (last bar in fig 2(c))- as explained in the sec 3.2 this approach do not apply directly to any of the other cameras.

The choice of prior is critical, and resolution is significantly reduced in the absence of an explicit slope model. For example, if the plenoptic camera samples include aliasing, the last row of figure 4 demonstrates that with a proper slope model we can super-resolve the plenoptic camera measurements, and the actual information encoded by the recorded plenoptic data is higher than that of the direct measurements.

The relative ranking of cameras also changes as a function of prior- while the plenoptic camera produced best results for the isotropic priors, a stereo camera achieves a higher resolution under the MOG prior. Our goal in the next section is to analytically evaluate the reconstruction accuracy of different cameras, and to understand how it is affected by the choice of prior.

4 Camera Evaluation

Given a light field prior we want to assess how well a light field x^0 can be recovered from a noisy projection $y = Tx^0 + n$, or how much the projection y nails down the set of possible light field interpretations. The uncertainty can be measured by the expected reconstruction error:

$$E(|W(x - x^0)|^2; T) = \int_x P(x|y; T)|W(x - x^0)|^2 \tag{7}$$

where $W = diag(w)$ is a diagonal matrix specifying how much we care about different light field entries, as discussed in sec 3.1. This measure should prefer distributions centered at the true solution, and whose variance around this solution is small as well (and thus, less likely to be shifted by noise).

To understand this measure, consider the 3 distributions in figure 5. The first distribution obtains a high reconstruction error since its peak is located away from the

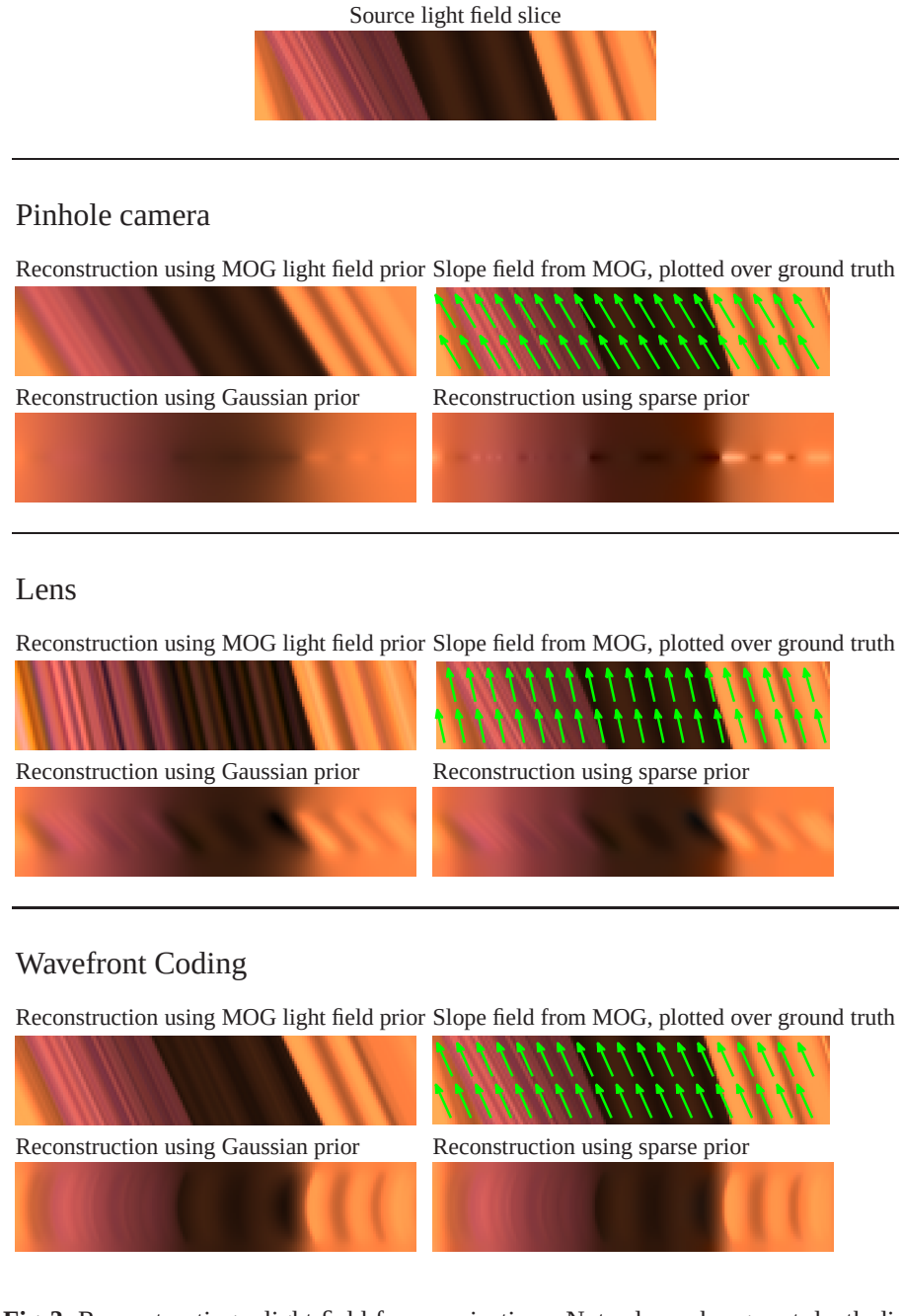


Fig. 3. Reconstructing a light field from projections. Note slope changes at depth discontinuities.

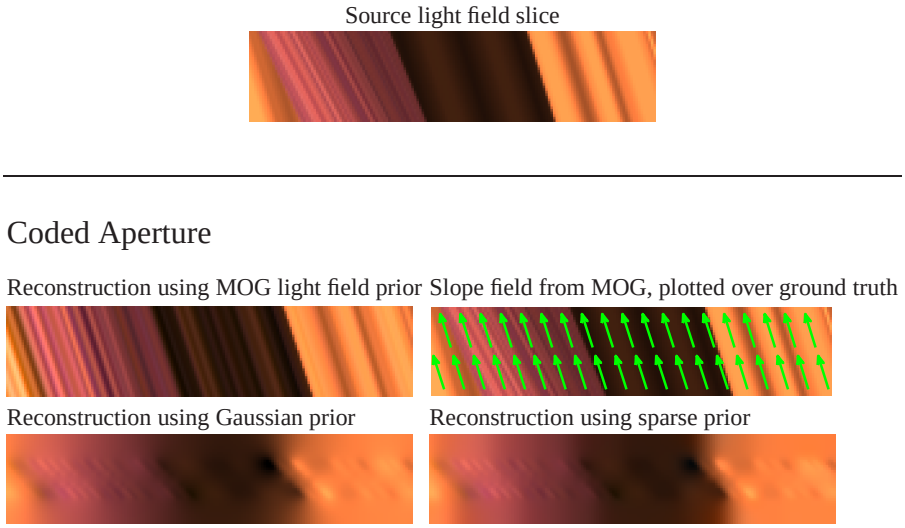
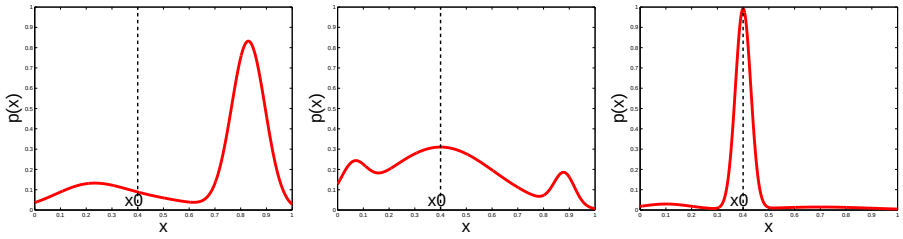


Fig. 4. Reconstructing a light field from projections, (continued). Note slope changes at depth discontinuities

original light field x^0. The second one is centered at the right solution, but the expected reconstruction error is still high due to the large variance around this solution. Such a high variation suggests that the projection does not nail down x^0 very firmly and the estimate can be easily shifted by noise. In contrast the third distribution achieves the smallest expected reconstruction error, being peaked and centered at the true solution.



$E(|x - x^0|^2) = 0.14$ $E(|x - x^0|^2) = 0.06$ $E(|x - x^0|^2) = 0.01$

Fig. 5. Uncertainty in estimation: The first two distributions will both lead to a high averaged error, while the third is picked at the true solution.

Uncertainty computation To simplify eq 7, recall that the average distance between x^0 and the elements of a Gaussian is the distance from the center, plus the variance:

$$E(|W(x - x^0)|^2|S;T) = |W(\mu_S - x^0)|^2 + \sum diag(W^2\Sigma_S) \tag{8}$$

In a mixture model, we need to weigh the contribution of each component by its overall volume:

$$E(|W(x - x^0)|^2;T) = \int_S P(S|y)E(|W(x - x^0)|^2|S;T) \tag{9}$$

Since the integral in eq 9 can not be computed explicitly, we evaluate an approximated uncertainty in the vicinity of the true solution, and we approximate eq 9 using a small set of slope field samples around the true slope interpretation. This is based on the assumption that for slope fields S which are very far from the true one, $P(y|S)$ is small and does not contribute much to the overall integral.

Finally, we use a set of typical light fields x_t^0 (generated using ray tracing) and evaluate the quality of a camera T as the expected squared error over these examples

$$E(T) = \sum_t E(|W(x - x_t^0)|^2;T) \tag{10}$$

Note that this solely measures information captured by the optics together with the prior, and omits the confounding effect of specific inference algorithms.

5 Tradeoffs in projection design

We can now study the reconstruction error of different designs and how it is affected by the light field prior.

Gaussian prior. We start by considering the generic isotropic Gaussian prior in eq 2. If the distribution of light fields x is Gaussian, we can integrate over x in eq 10 analytically to obtain: $E(T) = 2\sum diag(1/\eta^2 T^T T + \Psi_0^{-1})^{-1}$ Thus, we reach the classical

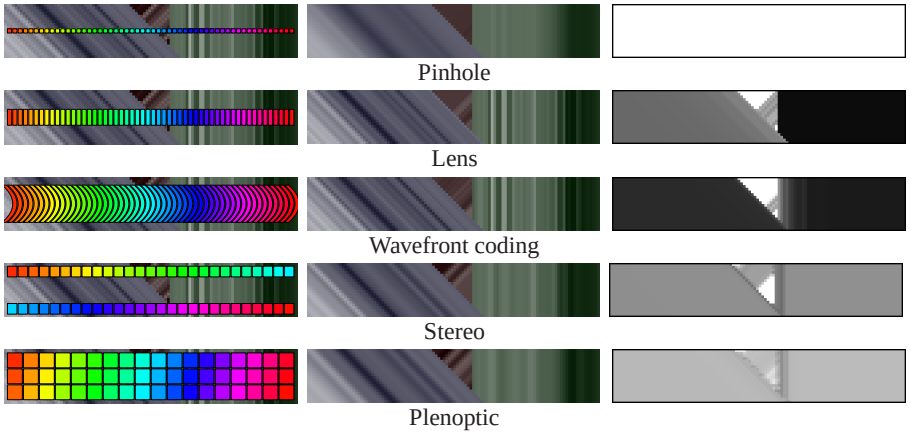


Fig. 6. Evaluating conditional uncertainty in light field estimate. Left: projection model. Middle: estimated light field. Right: variance in estimate (equal intensity scale used for all cameras). Note that while for visual clarity we plot perfect square samples, in our implementation samples were convolved with low pass filters to simulate realistic optics blurs.

principal components conclusion: to minimize the residual variance, T should measure the directions of maximal variance in Ψ_0. Since the prior is shift invariant, Ψ_0^{-1} is a convolution matrix, diagonal in the frequency domain, and the principal components are the lowest frequencies. Thus, an isotropic Gaussian prior agrees with the classical signal processing conclusion [33,34,35,36,37] - to sample the light field one should convolve with a low pass filter to meet the Nyquist limit and sample both the directional and spatial axis, as with a plenoptic camera configuration. (if the depth in the scene is bounded, fewer directional samples can be used [33]). This is also consistent with our empirical prediction, as for the Gaussian prior, the plenoptic camera indeed achieved the lowest error in fig 2(c). However, this sampling conclusion is conservative as the directional axis is more redundant than the spatial one. The source of the problem is the fact that second order statistics captured by a Gaussian distribution do not capture the high order dependencies of light fields.

Mixture of Gaussian light field prior. We now turn to the more realistic MOG prior introduced in sec 3.3. While the optimal projection under this prior cannot be predicted in closed-form, it can help us understand the major components influencing the performance of existing camera configurations. The score in eq 9 reveals two aspects which affect the quality of a camera- first, minimizing the variance Σ_S of each of the mixture components (i.e., the ability to reliably recover the light field given the true slope field), and second, the need to identify depth and make $P(S|y)$ peaked at the true slope field. Below, we elaborate on these two components.

5.1 Conditional light field estimation – known depth

Fig 6 shows light fields estimated by several cameras, assuming the true depth (and therefore slope field), was successfully estimated. We also display the variance of the estimated light field - the diagonal of Σ_S (eq 6).

In the right part of the light field, the lens reconstruction is sharp, since it averages rays emerging from a single object point. On the left, the lens reconstruction involves a higher uncertainty, since the lens averages light rays from multiple object points and blurs high frequencies. In contrast, integrating over a parabolic curve (wavefront coding) achieves low uncertainties for both slopes, since a parabola "covers" all slopes[2]. A pinhole also behaves identically at all depths, but it collects only a small amount of light and the uncertainty is high due to the small signal to noise ratio. Finally, the uncertainty increases in stereo and plenoptic cameras due to the smaller number of spatial samples.

The central region of the light field demonstrates the utility of multiple viewpoint in the presence of occlusion boundaries. Occluded parts which are not measured properly lead to higher variance. The variance in the occluded part is minimized by the plenoptic camera, the only one that spends measurements in this region of the light field.

Since we deal only with spatial resolution, our conclusions correspond to known imaging common sense, which is a good sanity check for our model. However, note that they cannot be derived from a naive Gaussian model, which emphasizes the need for a prior such as as our new mixture model.

5.2 Depth estimation

Light field reconstruction involves slope (depth) estimation. Indeed, the error in eq 9 also depends on the uncertainty about the slope field S. We need to make $P(S|y)$ peaked at the true slope field. Since the observation y is $Tx + n$, we want the distributions of projections Tx to be as distinguishable as possible for different slope fields S. One way to achieve this is to make the projections corresponding to different slope fields concentrated within different subspaces of the N-dimensional space. For example, a stereo camera yields a linear constraint on the projection- the $N/2$ samples from the first view should be a shifted version of the other $N/2$. The coded aperture camera also imposes linear constraints: certain frequencies of the defocused signals are zero, and the location of these zeros shifts with depth [2].

To test this, we measure the probability of the true slope field, $P(S|y)$, averaged over a set of test light fields (created with ray tracing). The stereo score is $< P(S|y) >= 0.95$ (where $< P(S|y) >= 1$ means perfect depth discrimination) compared to $< P(S|y) >= 0.84$ for coded aperture. This suggests that the disparity constraint of stereo better distributes the projections corresponding to different slope fields than the zero frequency subspace in coded aperture. On the other hand, while linear dependency among the elements of y helps us identify slopes, it means we are measuring less dimensions of x, and the variance in $P(x|y, S)$ is higher. For example, the y resulting from a plenoptic camera measurement lies in an N/k dimensional space (where k is the number of views), comparing to an $N/2$ dimensions of a stereo camera. The accuracy of the depth estimation in the plenoptic camera was increased to 0.98. This value is not significantly higher than stereo, while as demonstrated in figure 6,

[2] When depth is locally constant and the surface diffuse, we can map a light field integration curve into a classical Point Spread Function (PSF), by projecting it along the slope direction s. Projecting a parabola $\{(a, b)|b = a^2\}$ at direction s yields the PSF $psf(b) = |b - s/2|^{-0.5}$. That is, the PSF at different depths are equal up to spatial shift, which does not affect visual quality or noise sensitivity

the planoptic camera increases the variance in estimating x due to the loss of spatial resolution.

We can also use the averaged $P(S|y)$ score to quantitatively compare stereo with depth from defocus (DFD) - two lenses with the same center of projection, focused at two different depths. As predicted by [13], when the same physical size is used (stereo baseline shift doesn't exceed aperture width) both designs perform similarly, with DFD achieving $< P(S|y) >= 0.92$.

Our probabilistic treatment of depth estimation goes beyond linear subspace constraints. For example, the average slope estimation score of a lens was $< P(S|y) >= 0.74$, indicating that, while weaker than stereo, a single monocular image captured with a standard lens contains some depth-from-defocus information as well. This result cannot be derived using a disjoint-subspace argument, but if the full probability is considered, Occam's razor principle applies and the simpler explanation is preferred. To see why, suppose we are trying to distinguish between 2 constant slope explanation S_{focus} corresponding to the focus depth, and $S_{defocus}$ corresponding to one of the defocus depths. The set of images at a defocus depth (which includes images with low frequencies only) is a subset of the set of images at the focus depth (including both low and high frequency images). Thus, while a high frequency image can be explained only as an object at the focus depth, a low frequency image can be legally explained by both. However, since a probability sums to one, and since the set of defocus images occupies a smaller volume in the N-dimensional space, the defocus model assigns individual low frequency instances a higher probability.

Finally, a pinhole camera-projection just slices a row out of the light field, and this slice is invariant to the light field slope. The parabola filter of a wavefront coding lens is also designed to be invariant to depth. Indeed, for these two cameras, the evaluated distribution $P(S|y)$ in our model is uniform over slopes.

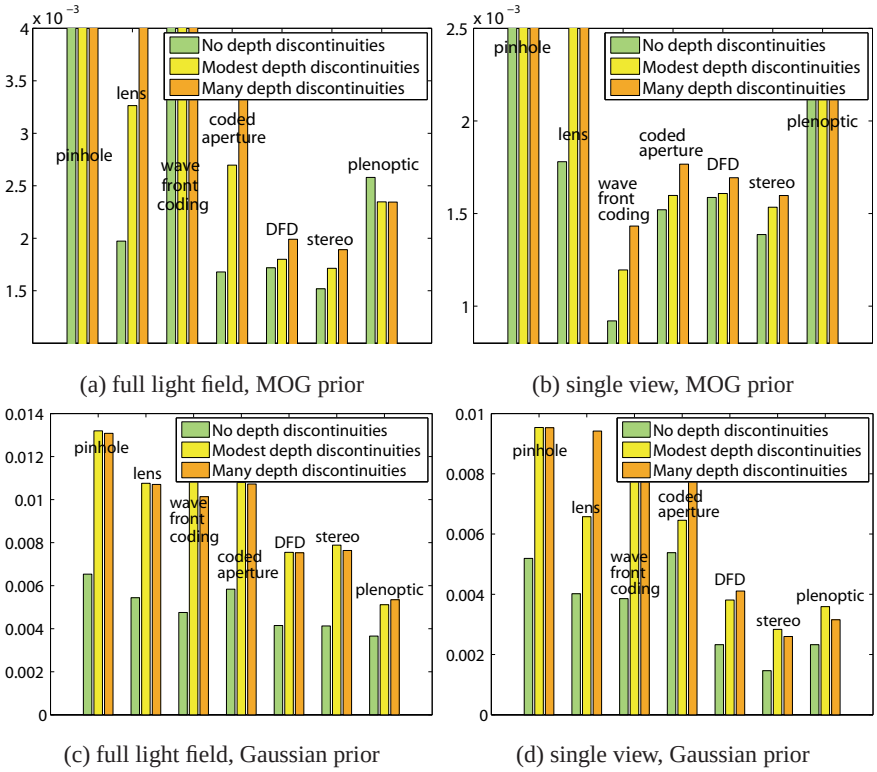
Again, these results are not fully surprising but they are obtained within a general framework that can qualitatively and quantitatively compare a variety of camera designs. While comparisons such as DFD vs. stereo have been conducted in the past [13], our framework encompasses a much broader family of cameras.

5.3 Light field estimation

In the previous section we gained intuition about the various parts of the expected error in eq 9. We now use the overall formula to evaluate existing cameras, using a set of diffuse light fields generated using ray tracing. Evaluated camera configurations include a pinhole camera, lens, stereo pair, depth-from-defocus (2 lenses focused at different depths), plenoptic camera, coded aperture cameras and a wavefront coding lens. Another advantage of our framework is that we can search for optimal parameters within each camera family, and our comparison is based on optimized parameters such as baseline length, aperture size and focus distance of the individual lens in a stereo pair, and various choices of codes for coded aperture cameras.

By changing the weights, W on light field entries in eq 7, we evaluate cameras for two different goals: (a) Capturing a full light field. (b) Achieving an all-focused image from a single view point (capturing a single row in the light field.)

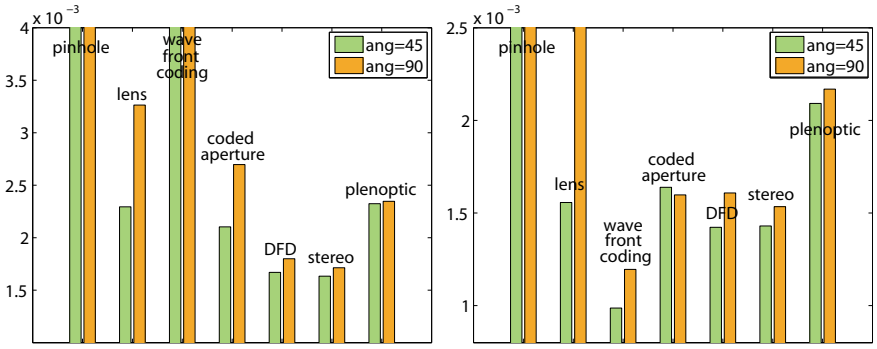
We consider both a Gaussian and our new mixture of Gaussians (MOG) prior. We consider different levels of depth complexity as characterized by the amount of dis-



(a) full light field, MOG prior (b) single view, MOG prior

(c) full light field, Gaussian prior (d) single view, Gaussian prior

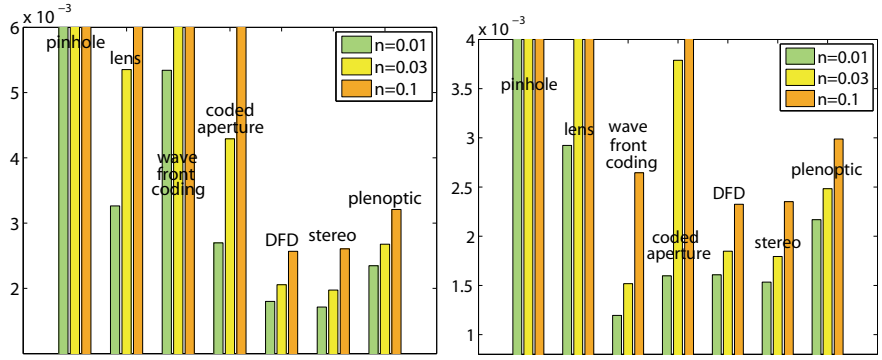
Fig. 7. Evaluating expected reconstruction error as a function of depth complexity.



(a) full light field, MOG prior (b) single view, MOG prior

Fig. 8. Evaluating expected reconstruction error as a function of slope range.

continuities. We use slopes between -45^o to 45^o and noise with standard deviation $\eta = 0.01$. Fig. 7(a-b) plot expected reconstruction error with our MOG prior, while



(a) full light field, MOG prior (b) single view, MOG prior

Fig. 9. Evaluating expected reconstruction error as a function of noise.

figs 7(c-d) use a generic isotropic Gaussian prior (note the different axis scale). In figure 8 we evaluate changes in the depth range (using light fields with modest amount of depth discontinuities and $\eta = 0.01$), and in figure 9 changes in the noise level (using light fields with modest amount of depth discontinuities, and slopes ranging between -45^o to 45^o).

Full light field reconstruction Fig. 7(a) shows full light field reconstruction with our MOG prior. In the presence of depth discontinues, lowest light field reconstruction is achieved with a stereo camera.

While a plenoptic camera improves depth information our comparison suggests it may not pay for the large spatial resolution loss. Yet, as discussed in sec 5.1 a plenoptic camera offers an advantage in the presence of complex occlusion boundaries.

For planar scenes (in which estimating depth is easy) the coded aperture surpasses stereo, since spatial resolution is doubled and the irregular sampling of light rays can avoid high frequencies loss due to defocus blur.

While the performance of all cameras decreases when the depth complexity increases, a lens and coded aperture are much more sensitive than others.

While the depth discrimination of DFD is similar to that of stereo (as discussed in sec 5.2), its overall reconstruction error is slightly higher since the wide apertures blur high frequencies.

The relative ranking in figs 7(a,c) agrees with the empirical prediction in figure 2(c). Note, however, that while figs 7(a,c) measure inherent optics information, fig 2(c) folds-in inference errors as well.

Single-image reconstruction When addressing the single row reconstruction goal (fig 7(b)) one still has to account for issues like defocus, depth of field, signal to noise ratio and spatial resolution. Thus, a pinhole camera (recording this single row alone) is not ideal, and there is an advantage for wide aperture configurations collecting more light (recording multiple light field rows) despite not being invariant to depth.

The parabola filter (wavefront coding) does not capture depth information and thus performs very poorly for the light field estimation goal. However, the evaluation in fig 7(b) suggests that for the goal of recovering a single light field row, this filter outperforms all other cameras. The reason is that since the filter is invariant to slope, a single central light field row can be recovered without knowledge of depth. For this central row, it actually achieves high signal to noise ratios for all depths, as demonstrated in figure 6. To validate this observation, we have searched over a large set of lens curvatures, or light field integration curves, parameterized as splines fitted to 6 key points. This family includes both slope sensitive curves (in the spirit of [8] or a coded aperture), which identify slope and use it in the estimation, and slope invariant curves (like the parabola [7]), which estimate the central row regardless of slope. Our results show that, for the goal of recovering a single light field row, the wavefront-coding parabola outperforms all other configurations. This extends the arguments in previous wavefront coding publications which were derived using optics reasoning and focus on depth-invariant approaches.

5.4 Plenoptic sampling: signal processing and Bayesian estimation

As another way to compare the conclusions derived by classical signal processing approaches with the ones derived from our new MOG light filed prior, we follow [33] and ask: suppose we use a camera with a fixed N pixels resolution, how many different views (N pixels each) do we actually need for a good 'virtual reality'?

Figure 10 plots the expected reconstruction error as a function of the number of views for both MOG and naive Gaussian priors. While a Gaussian prior requires dense sampling, the MOG error is quite low after 2-3 views (such conclusions depend on depth complexity and the range of views we wish to capture). For comparison, we also mark on the graph the significantly larger views number imposed by a Nyquist limit analysis, like [33]. Note that to simulate a realistic camera, our directional axis samples are aliased. This is slightly different from [33] which blur the directional axis in order to eliminate frequencies above the Nyquist limit.

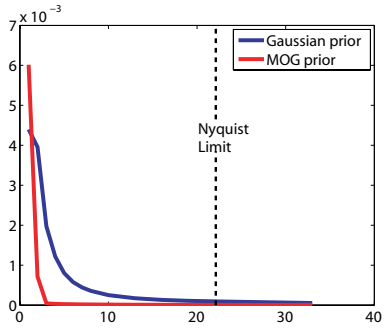


Fig. 10. Reconstruction error as a function number of views.

6 Discussion

The growing variety of computational camera designs calls for a unified way to analyze their tradeoffs. We show that all cameras can be analytically modeled by a linear mapping of light rays to sensor elements. Thus, interpreting sensor measurements is the Bayesian inference problem of inverting the ray mapping. We show that a proper light fields prior is critical for the successes of camera decoding. We analyze the limitations of traditional band-pass assumptions and suggest that a prior which explicitly accounts for the elongated light field structure can significantly reduce sampling requirements.

Our Bayesian framework estimates both depth and image information, accounting for noise and decoding uncertainty. This provides a tool to compare computational cameras on a common baseline and provides a foundation for computational imaging. We conclude that for diffuse scenes, the wavefront coding cubic lens (and the parabola light field curve) is the optimal way to capture a scene from a single view point. For capturing a full light field, a stereo camera outperformed other known configurations.

We have focused on providing a common ground for all designs, at the cost of simplifying optical and decoding aspects. This differs from traditional optics optimization tools such as Zemax [32] that provide fine-grain comparisons between subtly-different designs (e.g. what if this spherical lens element is replaced by an aspherical one?). In contrast, we are interested in the comparison between families of imaging designs (e.g. stereo vs. plenoptic vs. coded aperture). We concentrate on measuring inherent information captured by the optics, and do not evaluate camera-specific decoding algorithms.

The conclusions from our analysis are well connected to reality. For example, it can predict the expected tradeoffs (which can not be derived using more naive light field models) between aperture size, noise and spatial resolution discussed in sec 5.1. It justifies the exact wavefront coding lens design derived using optics tools, and confirms the prediction of [13] relating stereo to depth from defocus.

Analytic camera evaluation tools may also permit the study of unexplored camera designs. One might develop new cameras by searching for linear projections that yield optimal light field inference, subject to physical implementation constraints. While the camera score is a very non-convex function of its physical characteristics, defining camera evaluation functions opens up these research directions.

7 Appendix

This appendix extends section 3.3 to provide details on the slope field (depth) inference under our MOG light field prior.

Given a camera T and an observation y our goal is to infer a MAP estimation of x. The probability of a light field explanation $p(x|y)$ is defined as:

$$P(x|y;T) = \int_S P(S|y;T)P(x|y,S;T) \tag{11}$$

however, the integral in eq 11 is intractable. Our strategy was to compute an approximated MAP estimate for the slope field S, and conditioning on this estimated slope field, solve for the MAP light field.

To compute an approximated MAP estimate for the slope field, we break the light field into small overlapping windows $\{w\}$ along the spatial axis, and pick y_{S_w} - the m most central entries of y according to the slope orientation, as illustrated in fig 11. We can then ask locally what is $P(y_{S_w}|S_w)$, or how well are the measurements y_{S_w} explained by the S_w slope field window interpretation. For example, if we use a stereo camera, the local y_{S_w} measurements should satisfy the disparity shift constraints imposed by S_w. We approximate the slope score as a product over local windows, that is, we look for a slope field S maximizing:

$$P(S|y) \approx \prod_w P(S_w|y_{S_w}) \tag{12}$$

If we consider sufficiently small light field windows, we can reasonably cover the set of slope field interpretations with a discrete list $\{\mathbf{S^1}, ..., \mathbf{S^K}\}$. The list $\{\mathbf{S^1}, ..., \mathbf{S^K}\}$ we use includes constant slope field windows and slope fields windows with one depth discontinuity. We approximate the $P(\mathbf{S^i}|y_{\mathbf{S^i}})$ integral with a discrete sum:

$$P(\mathbf{S^i}|y_{\mathbf{S^i}}) \approx \frac{P(\mathbf{S^i})P(y_{\mathbf{S^i}}|\mathbf{S^i})}{\frac{1}{K}\sum_{k=1}^{K} P(\mathbf{S^k})P(y_{\mathbf{S^i}}|\mathbf{S^k})} \tag{13}$$

We optimize eq 12 using Belief Propagation (enforcing the slope fields in neighboring windows to agree). The exact window size poses a tradeoff- smaller windows will increase the efficiency of the computation but also decrease the robustness of the approximation.

We note that this algorithm is a generalization of other camera decoding algorithms. For example if the number of central y entries m is decreased to two pixels we achieve the classical MRF stereo matching. The coded aperture used a similar framework as well, except that only constant depth interpretations were considered in each window, and $P(S_w|y_{S_w})$ were approximated using maximum likelihood.

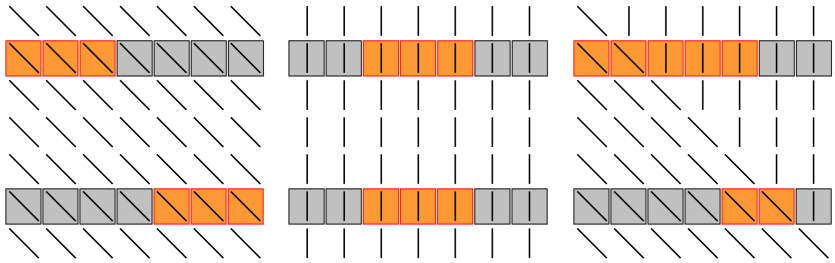


Fig. 11. Small slope field windows and the central y samples (highlighted in red), for a stereo camera

References

1. Fenimore, E., Cannon, T.: Coded aperture imaging with uniformly redundant rays. Applied Optics (1978) 1
2. Levin, A., Fergus, R., Durand, F., Freeman, W.: Image and depth from a conventional camera with a coded aperture. SIGGRAPH (2007) 1, 4, 7, 13
3. Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled photography: Mask-enhanced cameras for heterodyned light fields and coded aperture refocusing. SIGGRAPH (2007) 1, 4
4. Adelson, E.H., Wang, J.Y.A.: Single lens stereo with a plenoptic camera. IEEE PAMI (1992) 1, 4
5. Ng, R., Levoy, M., Bredif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Stanford U. Tech Rep CSTR 2005-02 (2005) 1, 4

6. Georgeiv, T., Zheng, K.C., Curless, B., Salesin, D., Nayar, S., Intwala, C.: Spatio-angular resolution tradeoffs in integral photography. In: EGSR. (2006) 1
7. Bradburn, S., Dowski, E., Cathey, W.: Realizations of focus invariance in optical-digital systems with wavefront coding. Applied optics **36** (1997) 9157–9166 1, 4, 17
8. Dowski, E., Cathey, W.: Single-lens single-image incoherent passive-ranging systems. App Opt (1994) 1, 17
9. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Intl. J. Computer Vision **47** (2002) 7–42 1, 4, 7
10. Levoy, M., Hanrahan, P.M.: Light field rendering. In: SIGGRAPH. (1996) 1, 2
11. Wilburn, B., Joshi, N., Vaish, V., Talvala, E., Antunez, E., Barth, A., Adams, A., Levoy, M., Horowitz, M.: High performance imaging using large camera arrays. SIGGRAPH (2005) 1
12. Zomet, A., Feldman, D., Peleg, S., Weinshall, D.: Mosaicing new views: The crossed-slits projection. PAMI (2003) 1
13. Schechner, Y., Kiryati, N.: Depth from defocus vs. stereo: How different really are they. IJCV (2000) 1, 14, 18
14. Pentland, A.P.: A new sense for depth of field. IEEE Trans. Pattern Anal. Mach. Intell. **9** (1987) 523–531 1
15. Grossmann, P.: Depth from focus. Pattern Recognition Letters **5** (1987) 63–69 1
16. Hasinoff, S.W., Kutulakos, K.N.: Confocal stereo. In: European Conference on Computer Vision. (2006) I: 620–634 1
17. Favaro, P., Mennucci, A., Soatto, S.: Observing shape from defocused images. Int. J. Comput. Vision **52** (2003) 25–43 1
18. Chaudhuri, S., Rajagopalan, A.: Depth from defocus: A real aperture imaging approach. Springer-Verlag, New York (1999) 1
19. Hiura, S., Matsuyama, T.: Depth measurement by the multi-focus camera. In: CVPR, IEEE Computer Society (1998) 953–961 1
20. Farid, H., Simoncelli, E.P.: Range estimation by optical differentiation. Journal of the Optical Society of America **15** (1998) 1777–1786 1
21. Greengard, A., Schechner, Y., Piestun, R.: Depth from diffracted rotation. Optics Letters **31** (2006) 181–183 1
22. Zhang, L., Nayar, S.K.: Projection Defocus Analysis for Scene Capture and Image Display. ACM Trans. on Graphics (also Proc. of ACM SIGGRAPH) (2006) 1
23. Hasinoff07, S., Kutulakos, K.: A layer-based restoration framework for variable-aperture photography. In: ICCV. (2007) 1
24. Green, P., Sun, W., Matusik, W., Durand, F.: Multi-aperture photography. SIGGRAPH (2007) 1
25. Moreno-Noguer, F., Belhumeur, P., Nayar, S.: Active Refocusing of Images and Videos. ACM Trans. on Graphics (also Proc. of ACM SIGGRAPH) (2007) 1
26. Kuthirummal, S., Nayar, S.K.: Multiview Radial Catadioptric Imaging for Scene Capture. ACM Trans. on Graphics (also Proc. of ACM SIGGRAPH) (2006) 1
27. Zomet, A., Nayar, S.: Lensless Imaging with a Controllable Aperture. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2006) 1
28. Nayar, S.K., Branzoi, V., Boult, T.E.: Programmable Imaging: Towards a Flexible Camera. International Journal on Computer Vision (2006) 1
29. Takhar, D., Laska, J., Wakin, M.B., Duarte, M.F., Baron, D., Sarvotham, S., Kelly, K.F., Baraniuk, R.G.: A new compressive imaging camera architecture using optical-domain compression. In: Computational Imaging IV at SPIE Electronic Imaging. (2006) 1
30. Fergus, R., Torralba, A., Freeman, W.T.: Random lens imaging. MIT CSAIL TR 2006-058 (2006) 1
31. Goodman, J.W.: Introduction to Fourier Optics. McGraw-Hill Book Company (1968) 2, 3

32. Zemax www.zemax.com. 2, 3, 18
33. Chai, J., Tong, X., Chan, S., Shum, H.: Plenoptic sampling. SIGGRAPH (2000) 2, 3, 4, 5, 8, 12, 17
34. Isaksen, A., McMillan, L., Gortler, S.J.: Dynamically reparameterized light fields. In: SIGGRAPH. (2000) 2, 3, 4, 5, 8, 12
35. Zwicker, M., Matusik, W., Durand, F., Pfister, H.: Antialiasing for automultiscopic displays. In: EGSR. (2006) 2, 3, 4, 5, 8, 12
36. Ng, R.: Fourier slice photography. SIGGRAPH (2005) 2, 3, 4, 5, 8, 12
37. Stewart, J., Yu, J., Gortler, S., McMillan, L.: A new reconstruction filter for undersampled light fields. In: Eurographics Rendering Workshop. (2003) 2, 3, 5, 8, 12
38. Seitz, S., Kim, J.: The space of all stereo images. In: ICCV. (2001) 3
39. Grossberg, M., Nayar, S.K.: The raxel imaging model and ray-based calibration. IJCV (2005) 3
40. Yu, J., McMillan, L.: General linear cameras. In: ECCV. (2004) 3
41. Adams, A., Levoy, M.: General linear cameras with finite aperture. In: EGSR. (2007) 3
42. Kak, A.C., Slaney, M.: Principles of Computerized Tomographic Imaging. 3
43. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE PAMI (2002) 3
44. Donoho, D.: Compressed sensing. Tech Rep (2004) 3, 8
45. Weiss, Y., Chang, H.S., Freeman, W.T.: Learning compressed sensing. In: Allerton. (2007) 3
46. Roth, S., Black, M.J.: Fields of experts: A framework for learning image priors. In: CVPR. (2005) 6, 8