# Image and Depth from a Conventional Camera with a Coded Aperture

Anat Levin Rob Fergus Frédo Durand William T. Freeman

Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory


**Figure 1: Left:** *Image captured using our coded aperture.* **Center:** *Top, closeup of captured image. Bottom, closeup of recovered sharp image.* **Right:** *Recovered depth map with color indicating depth from camera (cm) (in this this case, without user intervention).*

## Abstract

A conventional camera captures blurred versions of scene information away from the plane of focus. Camera systems have been proposed that allow for recording all-focus images, or for extracting depth, but to record both simultaneously has required more extensive hardware and reduced spatial resolution. We propose a simple modification to a conventional camera that allows for the simultaneous recovery of both (a) high resolution image information and (b) depth information adequate for semi-automatic extraction of a layered depth representation of the image.

Our modification is to insert a patterned occluder within the aperture of the camera lens, creating a coded aperture. We introduce a criterion for depth discriminability which we use to design the preferred aperture pattern. Using a statistical model of images, we can recover both depth information and an all-focus image from single photographs taken with the modified camera. A layered depth map is then extracted, requiring user-drawn strokes to clarify layer assignments in some cases. The resulting sharp image and layered depth map can be combined for various photographic applications, including automatic scene segmentation, post-exposure refocusing, or re-rendering of the scene from an alternate viewpoint.

**Keywords:** Computational Photography, Coded Imaging, Depth of field, Range estimation, Image statistics, Deblurring

**ACM Reference Format**
Levin, A., Fergus, R., Durand, F., Freeman, W. 2007. Image and Depth from a Conventional Camera with a Coded Aperture. *ACM Trans. Graph. 26*, 3, Article 70 (July 2007), 9 pages. DOI = 10.1145/1239451.1239521 http://doi.acm.org/10.1145/1239451.1239521.



## 1 Introduction

Traditional photography captures only a 2-dimensional projection of our 3-dimensional world. Most modifications to recover depth require multiple images or active methods with extra apparatus such as light emitters. In this work, with only minimal change from a conventional camera system, we seek to retrieve coarse depth information together with a normal high resolution RGB image. Our solution uses a single image capture, and a small modification to a traditional lens – a simple piece of cardboard suffices – together with occasional user assistance. This system allows photographers to capture images the same way they always have, but provides coarse depth information as a bonus, allowing refocusing (or an extended depth of field) and depth-based image editing.

Our approach is an example of *computational photography* where an optical element alters the incident light array so that the image captured by the sensor is not the final desired image but is *coded* to facilitate the extraction of information. More precisely, we build on ideas from coded aperture imaging [Fenimore and Cannon 1978] and wavefront coding [Cathey and Dowski 1995; Dowski and Cathey 1994] and modify the defocus produced by a lens to enable both the extraction of depth information *and* the retrieval of a standard image. Our contribution contrasts with other approaches in this regard - they recover either the image or the depth but not both from a single image. The principle of our approach is to control the effect of defocus so that we can both estimate the amount of defocus easily – and hence infer distance information – while at the same time making it possible to compensate for at least part of the defocus to create artifact-free images.

**Principle** To understand how we can control and exploit defocus, consider Figure 2 which illustrates a simplified thin lens model that maps light rays from the scene onto the sensor. When an object is placed at the focus distance $D$, all the rays from a point in the scene will converge to a single sensor point and the output image will appear sharp. Rays from an object at a distance $D_k$, away from the focus distance, land on multiple sensor points resulting in a blurred image. The pattern of this blur is given by the aperture cross section of the lens and is often called a circle of confusion. The amount of defocus, characterized by the blur radius, depends on the distance of the object from the focus plane.

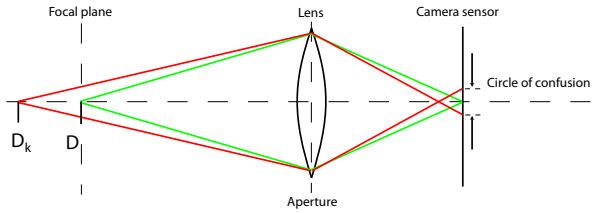



**Figure 2:** *A 2D thin lens model. At the plane of focus, a distance D from the lens, light rays (shown in green) emanating from a point are focused to a point on the camera sensor. Rays from a point at a distance $D_k$ (shown in red) no longer map to a point but rather to a region of the sensor, known as the circle of confusion. The pattern within this circle is determined by the aperture shape.*

For a simple planar object at distance $D_k$, the imaging process can be modeled as a convolution:

$$y = f_k * x \tag{1}$$

where $y$ is the observed image, $x$ is the true sharp image and the blur filter $f_k$ is a scaled version of the aperture shape (potentially convolved with the diffraction pattern). Figure 3(a) shows the pattern of blur from a conventional lens, the pentagonal disk shape being formed by the intersecting diaphragm blades. The defocus from such aperture does provide depth cues, e.g. [Pentland 1987], but they are challenging to exploit because it is difficult to precisely estimate the amount of blur and it requires multiple images.

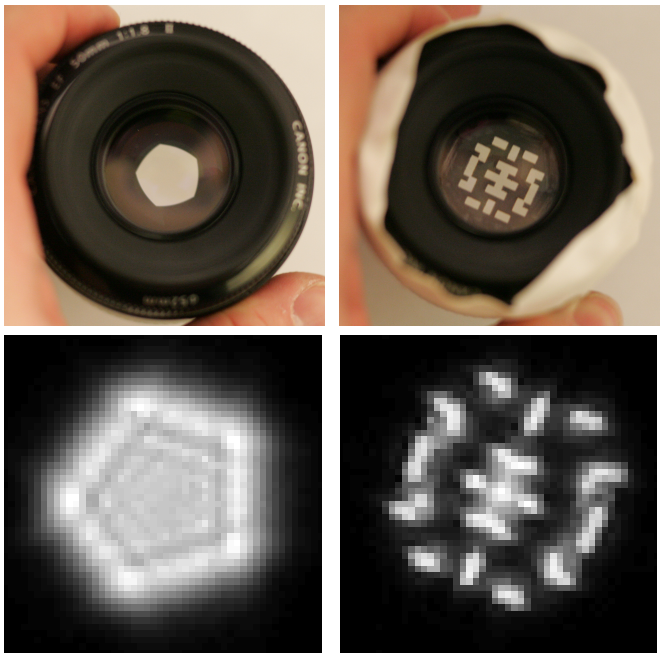
In this paper we explore what happens if patterns are deliberately introduced into the aperture, as illustrated in Figure 3(b). As before, the captured image will still be blurred as a function of depth with the blur being a scaled version of the aperture shape, but the aperture filter can be designed to discriminate between different depths.

Revisiting the image formation Eqn. 1 and assuming the aperture shape is known and fixed, only a single unknown parameter relates the blurred image $y$ to its sharp version $x$ – the scale of the blur filter. However in real scenes the depth is rarely constant throughout. Instead the scale of the blur in the image $y$, while locally constant, will vary over its extent. So the challenge is to recover not just a single blur scale but a map of it over the image. If this can be reliably recovered it would have great practical utility. First, the depth of the scene can be directly computed. Second, we can decode the captured image $y$. That is, invert $f_k$ and so recover a fully sharp image $x$. Hence our approach promises the recovery of both a depth map and a sharp image from the single blurry image $y$. In this paper we explore how the scale map of the blur may be recovered from the captured image $y$, designing aperture filters which are highly sensitive to depth variations.

The above discussion only takes into account geometric optics. A more comprehensive treatment must include wave effects, and in particular diffraction. The diffraction caused by an aperture is the Fourier power spectrum of its cross section. This means that the defocus blurring kernel is the convolution of the scaled aperture shape with its own power spectrum. For objects in focus, diffraction dominates, but for defocused areas, the shape of the aperture is most important. Thus, the analysis of defocus usually relies on geometric optics. While our theoretical derivation is based on geometric optics, in practice we account for diffraction by calibrating the blur kernel from real data.

### 1.1 Related work

Depth estimation using optical methods is an active area of research which can be divided into two main approaches: active and passive.



(a) Conventional (b) Coded

**Figure 3: Left***: Top, a standard Canon 50mm $f/1.8$ lens with the aperture partially closed. Bottom, the resulting blur pattern. The intersecting aperture blades give the pentagonal shape, while the small ripples are due to diffraction.* **Right***: Top, the same model of lens but with our filter inserted into the aperture. Bottom, the resulting blur pattern, which allows recovery of both image and depth.*

Active methods include laser scanning [Axelsson 1999] and structured light methods [Nayar et al. 1995; Zhang and Nayar 2006]. While these approaches can produce high quality depth estimates, they involve additional illumination sources. In contrast, passive approaches aim to capture the world without such additional intervention, the 3D information being recovered by the analysis of changes in viewpoint or focus.

Multiple viewpoints may be obtained by capturing multiple images, as in stereo [Scharstein and Szeliski 2002]. Multiple viewpoints can also be collected in a single image using a plenoptic camera [Adelson and Wang 1992; Ng et al. 2005; Georgiev et al. 2006; Levoy et al. 2006] but at the price of a significant loss in the spatial resolution of the image. The second class of passive depth acquisition techniques are depth from focus and depth from defocus techniques [Pentland 1987; Grossmann 1987; Hasinoff and Kutulakos 2006; Favaro et al. 2003; Chaudhuri and Rajagopalan 1999], which involve capturing multiple images of the world from a single viewpoint using multiple focus settings. Depth is inferred from the analysis of changes in defocus. Some approaches to depth from defocus also make usage of optical masks to improve depth discrimination [Hiura and Matsuyama 1998; Farid and Simoncelli 1998; Greengard et al. 2006], although these approaches still require multiple images. Many of these depth from defocus methods have only been tested on highly textured images, unlike the conventional real photographs considered in this paper. Additionally, many of these methods have difficulty in accurately locating occlusion boundaries.

While depth acquisition techniques utilizing multiple images can potentially produce better depth estimates than our approach, they

are complicated by the need to capture multiple images, making them impractical in most personal photography settings. In this work our goal is to infer depth and an image from a single shot, without additional user requirements and without loss of image quality. There have been some previous attempts to use optical masks to recover depth from a single image, but none of these approaches demonstrated the reconstruction of a high quality image as well. Dowski and Cathey [1994] use a phase plate designed to be highly sensitive to depth variations but the image cannot be recovered. Other approaches like [Lai et al. 1992] demonstrate results only on synthetic bar images, and [Jones and Lamb 1993] presents only 1D plots of image rows.

The goal of the methods described above is to produce a depth image. Another approach, related to the other goal of our system, is to create an all-focus image, independent of depth. Wavefront coding [Cathey and Dowski 1995], deliberately defocuses the light rays using phase plates so that the defocus is the same at all depths, which then allows a single deconvolution to output an image with large depth of focus, but without allowing the simultaneous estimates of depth.

Coded aperture methods have been employed previously, notably in astronomy and medical imaging for X or gamma rays as a way of collecting more light, because traditional lenses cannot be used at these wavelengths. In most of these cases, all incoming light rays are parallel and hence blur scale estimation is not an issue, as the blur obtained is uniform over the image. These include generalizations of the pinhole camera called coded aperture imaging [Fenimore and Cannon 1978]. Similarly, Raskar et al [2006] applied coded exposure in the temporal domain for motion deblurring.

Our method exploits a statistical characterization of images to find the combination of depth-dependent blur and unblurred image that best explains the observed image. This is closely related to the blind deconvolution problem [Kundur and Hatzinakos 1996]. Despite recent progress in blind deconvolution using machine learning techniques, the problem is still highly challenging. Recent approaches assume the entire image is blurred uniformly [Fergus et al. 2006]. In [Levin 2006] the uniform blur assumption was somewhat relaxed, but restricting the discussion to a small family of 1D blurs.

### 1.2 Overview

The structure of the paper is as follows: Section 2 explains the design process for the coded filter and strategies for identifying the correct blur scale. In Section 3 we detail how the observed image may be deblurred to give a sharp image. Section 4 then explains how a depth map for the image can be recovered. We present our experimental results in Section 5, showing a calibrated lens capturing real scenes. Finally, we discuss the limitations of our approach and possible extensions.

Throughout the paper we will use lower case symbols to denote spatial domain signals with upper case corresponding to their frequency domain representations. Also, for a filter $f$, we define $C_f$ to be the corresponding convolution matrix (i.e. $C_f x \equiv f * x$). Similarly, $C_F$ will denote a convolution in the frequency domain (in this case, a diagonal matrix).

## 2 Aperture Filter Design

The two key requirements for an aperture filter are: (i) it is possible to reliably discriminate between the blurs that result from different scalings of the filter and (ii) the filter can be easily inverted so that the sharp image may be recovered. Given the huge space of possible filters, selecting the optimal filter under these two criteria is not a straightforward task. Before formally presenting our statistical approach, it will be useful to consider a simple example to build an intuition about the problem.

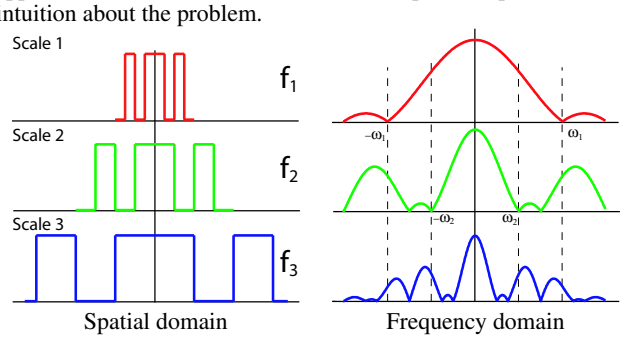



**Figure 4:** *A simple 1D example illustrating how the structure of zeros in the frequency domain shifts as a toy filter is scaled in the spatial domain.*

Figure 4 shows a 1D coded filter at 3 different scales, along with the corresponding Fourier transforms. The idea is to consider the structure of frequencies at which the Fourier transform of the filter is zero [Premaratne and Ko 1999]. For example, the filter $f_1$ (at scale 1) has a zero at $\omega_1$. This means that if the image $y$ was indeed blurred by $f_1$ then $Y(\omega_1) = 0$. Hence the zeros frequencies in the observed image can reveal the scale of the filter and hence its depth.

This argument can also be made in the spatial domain. If $Y(\omega_1) = 0$ it means that $y$ can no longer be an arbitrary $N$ dimensional vector ($N$ being the number of image pixels) as there are linear constraints it must satisfy. As the filter is scaled, the location of the zero frequencies shifts (e.g. moving from scale 1 to 2, the first zero moves from $\omega_1$ to $\omega_2$, see Figure 4). Hence each different scale defines a different linear subspace of possible blurry images. Given an $N$ dimensional input image, identifying the scale by which it was blurred (and thus identifying the object depth) reduces to identifying the subspace in which it lies.

While in theory identifying the zero frequencies or equivalently finding the correct subspace sounds straightforward, in practice the situation is more complex. First, noise in the imaging process means that no frequency will be exactly zeroed (thus the image $y$ will not exactly lie on any subspace). Second, zero frequencies in the observed image $y$ may just result from a zero frequency content in the original image signal $x$. This point is especially important since in order to account for depth variations, one would like to be able to make decisions based on small local image windows. These issues suggest that some aperture filters are better than others. For example, filters with zeros at low frequencies are likely to be more robust to noise than those with zeros at high frequencies, since a typical image has most of its energy at low frequencies. Also, if $\omega_1$ is a zero frequency of $f_1$, we want the filter at other scales $f_2, f_3$ etc. to have significant frequency content at $\omega_1$, so that we do not confuse the frequency responses.

Note that while a distinct pattern of zeros at each scale makes the depth identification easy, it makes inverting the filter hard since the deblurring procedure will be very sensitive to noise at these frequencies. To be able to retrieve depth information we must sacrifice some of the image content. However, if only a modest number of frequencies is sacrificed, the usage of image priors can reduce the noise sensitivity making it possible to reliably deblur kernels of moderate size ($\leq 15$ pixels). In this work, we mainly concentrate on optimizing the depth discrimination of the filter. This is the opposite focus of previous work such as [Raskar et al. 2006] where the coded filters were designed to have a very flat spectrum, eliminating zeros to make the deblurring as easy as possible.

To guide the design of the aperture filter, we introduce a statistical

model of real world images. Using this model we can compute the statistics of images blurred by a specific filter kernel. The model leads to a principled criterion for measuring the scale selectivity of a filter which we use as part of a random search over possible filters to pick a good one.

### 2.1 Statistical Model of Images

Real world images have statistics quite different from random matrices of white noise. One well known statistical property of images is that they have a sparse derivative distribution [Olshausen and Field 1996]. We impose this constraint during image reconstruction. However, in the filter design stage, to make our optimization tractable we assume that the distribution is Gaussian instead of the conventional heavy-tailed density. That is, our prior assumes the derivatives in the unobserved sharp image $x$ follow a Gaussian distribution with zero mean.

$$P(x) \propto \prod_{i,j} e^{-\frac{1}{2}\alpha\left((x(i,j)-x(i+1,j))^2+(x(i,j)-x(i,j+1))^2\right)} = \mathtt{N}(0,\Psi) \quad (2)$$

where $i, j$ are the pixel indices. $\Psi^{-1} = \alpha(C_{g_x}^T C_{g_x} + C_{g_y}^T C_{g_y})$, where $C_{g_x}, C_{g_y}$ are the convolution matrices corresponding to the derivative filters $g_x = [1 \;\; -1]$ and $g_y = [1 \;\; -1]^T$. Finally, the scalar $\alpha$ is set so the variance of the distribution matches the variance of derivatives in natural images ($\alpha = 250$ in our implementation). This image prior implies that the signal $x$ is smooth and its derivatives are often close to zero. The above prior can also be expressed in the frequency domain and, since derivatives are convolutions, the prior is diagonal in the frequency domain (if boundary effects are ignored):

$$P(X) \propto e^{-\frac{1}{2}\alpha X^T \bar{\Psi}^{-1} X} \text{ where } \bar{\Psi}^{-1} = \alpha\, diag(|G_x(\nu,\omega)|^2 + |G_y(\nu,\omega)|^2) \quad (3)$$

where $\nu, \omega$ are coordinates in the frequency domain. We observe a noisy blurred image which, assuming constant scene depth, is modeled as $y = f_k * x + n$. The noise in neighboring pixels is assumed to be independent, following a Gaussian model $n \sim \mathtt{N}(0, \eta^2 I)$ ($\eta = 0.005$ in our implementation). We denote $P_k(y)$ as the distribution of observed signals under a blur $f_k$ (that is, the distribution of images coming from objects at depth $D_k$). The blur $f_k$ linearly transforms the distribution of sharp images from Eqn. 2, so that $P_k(y)$ is also a Gaussian[1]: $P_k(y) \sim \mathtt{N}(0, \Sigma_k)$. The covariance matrix $\Sigma_k$ is a transformed version of the prior covariance, plus noise.

$$\Sigma_k = C_{f_k} \Psi C_{f_k}^T + \eta^2 I \xrightarrow[\text{transform}]{\text{Fourier}} \bar{\Sigma}_k = C_{F_k} \bar{\Psi} C_{F_k}^T + \eta^2 I \quad (4)$$

where transforming into the frequency domain makes the prior diagonal[2]. In the diagonal version, the distribution of the blurry image in the Fourier domain becomes:

$$P_k(Y) \propto exp(-\frac{1}{2} E_k(Y)) = exp(-\frac{1}{2} \sum_{\nu,\omega} |Y(\nu,\omega)|^2 / \sigma(\nu,\omega)) \quad (5)$$

where $\sigma(\nu,\omega)$ are the diagonal entries of $\bar{\Sigma}_k$:

$$\sigma(\nu,\omega) = |F_k(\nu,\omega)|^2 (\alpha |G_x(\nu,\omega)|^2 + \alpha |G_y(\nu,\omega)|^2)^{-1} + \eta^2 \quad (6)$$

Eqn. 6 represents a soft version of the zero frequencies test mentioned above. If the filter $f_k$ has a zero at frequency $(\nu,\omega)$ then $\sigma(\nu,\omega) = \eta^2$, typically a very small number. Thus, if the frequency content of the observed signal $Y(\nu,\omega)$ is significantly bigger than 0, the probability of $Y$ coming from the distribution $P_k$ is very low. In other words, if we find frequency content where the filter has a zero, it is unlikely that we have the correct scale of blur. We also note that the covariance at each frequency depends not only on $F(\nu,\omega)$ but also on our prior distribution, thus giving a smaller weight to higher frequencies which are less common in natural images.

[1] If $X, Y$ are random variables and $A$ a linear transformation with $X$ Gaussian and $Y = AX$, then $Cov(Y) = ACov(X)A^T$.

[2] This follows from (i) the Fourier transform of a convolution matrix is a diagonal matrix and (ii) all the matrices making up $\Sigma_k$ are either diagonal or convolution matrices.

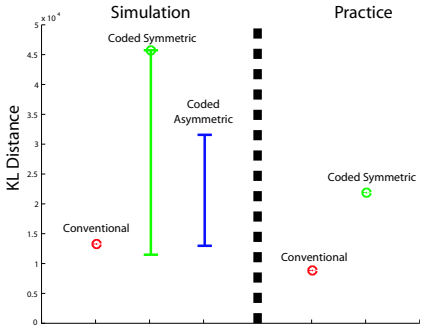



**Figure 5:** *A theoretical and practical comparison of conventional and coded apertures using the criterion of Eqn. 8. On the left side of the graph we plot the theoretical performance (KL distance – larger is better) for a conventional aperture (red), random symmetric coded filters (green error bar) and random asymmetric coded filters (blue error bar). On the right side of the graph we show the performance of the actual filters obtained in calibration, both of a conventional lens and a coded lens (see Figure 9). While the performance of the actual filters is lower than the theoretical prediction (probably due to high frequencies being lost in the imaging process), the coded filter still performs better than the conventional aperture.*

### 2.2 Filter Selection Criterion

The proposed model gives the likelihood of a blurry input image $y$ for a filter $f$ at a scale $k$. We now show how this may be used to measure the robustness of a particular aperture filter at identifying the true blur scale. Intuitively, if the blurry image distributions $P_{k_1}(y)$ and $P_{k_2}(y)$ at depths $k_1$ and $k_2$ are similar it will be hard to tell the depths apart. A classical measure of the distance between distributions is the Kullback–Leibler (KL) divergence:

$$D_{KL}(P_{k_1}(y), P_{k_2}(y)) = \int_y P_{k_1}(y)(logP_{k_1}(y) - logP_{k_2}(y))\, dy \quad (7)$$

A filter that maximizes this distance will have a typical blurry image at depth $k_1$ with a high likelihood under model $P_{k_1}(y)$ but a low likelihood under the model $P_{k_2}(y)$ for depth $k_2$. Using the frequency domain representation of our model (Eqns. 5 & 6) in Eqn. 7, the KL divergence reduces (up to a constant) to [3]

$$D_{KL}(P_{k_1}, P_{k_2}) = \sum_{\nu,\omega} \left( \frac{\sigma_{k_1}(\nu,\omega)}{\sigma_{k_2}(\nu,\omega)} - log\left( \frac{\sigma_{k_1}(\nu,\omega)}{\sigma_{k_2}(\nu,\omega)} \right) \right) \quad (8)$$

Eqn. 8 implies that the distance between the distributions of two different scales will be large when the ratio of their expected frequencies is high. This ratio may be maximized by having frequencies $\nu, \omega$ for which $F_{k_2}(\nu,\omega) = 0$ and $F_{k_1}(\nu,\omega)$ is large. This reflects the intuitions discussed earlier, that the zeros of the filter are useful for discriminating between different scales. For a zero in one scale to be particularly discriminative, other scales should maintain significant signal content in the same frequency. Also, the fact that $\sigma_k(\nu,\omega)$ weights the filter frequency content by the image prior (see Eqn. 6), indicates that zeros are more discriminative in lower frequencies, in which the original image is expected to have significant content.

### 2.3 Filter Search

Having introduced a principled criterion for evaluating a particular filter, we address the problem of searching for the optimal filter

[3] Since the probabilities are Gaussians, their log is quadratic, and hence the averaged log is the variance.

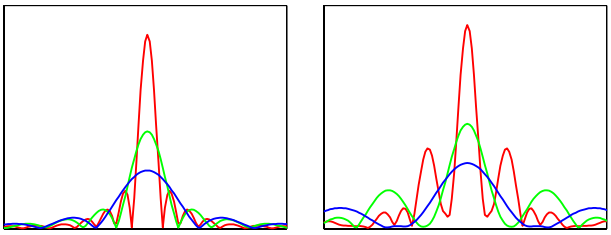



**Figure 6:** *The Fourier transforms of a 1D slide through the blur pattern from conventional and coded lenses at 3 different scales*

shape. When selecting a filter, a number of practical constraints should to be taken into account. First, the filter should be binary since non-binary filters are hard to construct accurately. Second, we should be able to cut the filter from a single piece of material, without having floating particles in the center. Third, to avoid excessive radial distortion (as explained in section 5), we avoid using the full aperture. Finally, diffraction imposes a minimum size on the holes in the filter.

Balancing these considerations, we confined our search to binary $13 \times 13$ patterns with $1\text{mm}^2$ holes. We randomly sampled a large number of $13 \times 13$ patterns. For each pattern, 8 different scales were considered, varying between 5 and 15 pixels in width. The random pattern was scored according to the minimum KL-divergence between the distributions of any two scales.

Figure 5 plots KL-divergence scores for the randomly generated filters, distinguishing between two classes of patterns – symmetric and asymmetric. Our observation was that symmetric patterns produce higher KL-divergence scores compared to asymmetric patterns. Examining the frequency structure of asymmetric filters we observed that such filters have few zero frequencies. By contrast, symmetric filters tend to produce a richer zeros structure. The symmetric pattern with the best score is shown in Figure 3(b). For comparison we also plotted the KL-divergence score for a conventional aperture. Also plotted in Figure 5 are the KL scores for actual filters obtained by calibrating a coded aperture lens and a conventional lens.

In Figure 6 we plot a 1D slices of the Fourier transform of both the best performing pattern and a conventional aperture at three different scales. In the case of the coded pattern each scale has a quite different frequency response, in particular their zeros occur at distinct frequencies. On the other hand, for the conventional aperture the zeros in different scales overlap heavily, making it hard to distinguish between them.

## 3 Deblurring

Having identified the correct blur scale of an observed image $y$, the next objective is to remove the blur, reconstructing the original sharp image $x$. This task is known as *deblurring* or *deconvolution*. Under our probabilistic model

$$P_k(x|y) \propto exp(-(\frac{1}{\eta^2}|C_{f_k}x-y|^2+\alpha|C_{g_x}x|^2+\alpha|C_{g_y}x|^2)) \quad (9)$$

The deblurring problem can thus be posed as finding the maximum likelihood explanation for $y$, $x^* = argmax\, P_k(x|y)$. For a Gaussian distribution, this reduces to a least squares optimization problem

$$x^* = argmin \frac{1}{\eta^2}|C_{f_k}x-y|^2+\alpha|C_{g_x}x|^2+\alpha|C_{g_y}x|^2 \quad (10)$$

By minimizing Eqn. 10 we search for the $x$ minimizing the reconstruction error $|C_{f_k}x-y|^2$, with the prior preferring $x$ to be as smooth as possible.

We note that the optimal solution to Eqn. 10 can be found by solving a sparse set of linear equations: $Ax = b$ for

$$A = \frac{1}{\eta^2}C_{f_k}^T C_{f_k}+\alpha C_{g_x}^T C_{g_x}+\alpha C_{g_y}^T C_{g_y} \qquad b = \frac{1}{\eta^2}C_{f_k}^T y \quad (11)$$

Eqn. 11 can be solved in the frequency domain in a few seconds for megapixel sized image. While this approach does produce wrap-around artifacts along the image boundaries, these are usually unimportant in large images.

Deblurring with a Gaussian prior on image derivatives is simple and efficient, but tends to over-smooth the result. To produce sharper decoded images, a stronger natural image prior is required, and a sparse derivatives prior was used. Thus, to solve for $x$ we minimize

$$|C_{f_k}x-y|+\sum_{ij}\rho(x(i,j)-x(i+1,j))+\rho(x(i,j)-x(i,j+1)) \quad (12)$$

where $\rho$ is a heavy-tailed function, in our implementation $\rho(z) = |z|^{0.8}$. While a Gaussian prior prefers to distribute derivatives equally over the image, a sparse prior opts to concentrate derivatives at a small number of pixels, leaving the majority of image pixels constant. This produces sharper edges, reduces noise and helps to remove unwanted image artifacts such as ringing. The drawback of a sparse prior is that the optimization problem is no longer a simple least squares one, and cannot be minimized in closed form (in fact, the optimization is no longer convex). To optimize this, we use an iterative reweighted least squares process e.g. [Levin and Weiss To appear] which poses the optimization as a sequence of least squares problems while the weight of each derivative is updated based on the previous iteration solution. The re-weighting means that Eqn. 11 cannot be solved in the frequency domain, so we are forced to work in the spatial domain using the Conjugate Gradient algorithm e.g. [Barrett et al. 1994]. The bottleneck in each iteration of this algorithm is the multiplication of each residual vector by the matrix $A$. Luckily the form of $A$ (Eqn. 11) enables this to be performed efficiently as a concatenation of convolution operations. However, this procedure still takes around 1 hour on a 2.4Ghz CPU for a 2 megapixel image. Our sparse deblurring code is available on the project webpage: `http://groups.csail.mit.edu/graphics/CodedAperture`.

Figure 7 demonstrates the difference between the reconstructions obtained with a Gaussian prior and a sparse prior. While the sparse prior produces a sharper image, both approaches produce better results than the classical Richardson-Lucy deconvolution scheme.

### 3.1 Blur Scale Identification

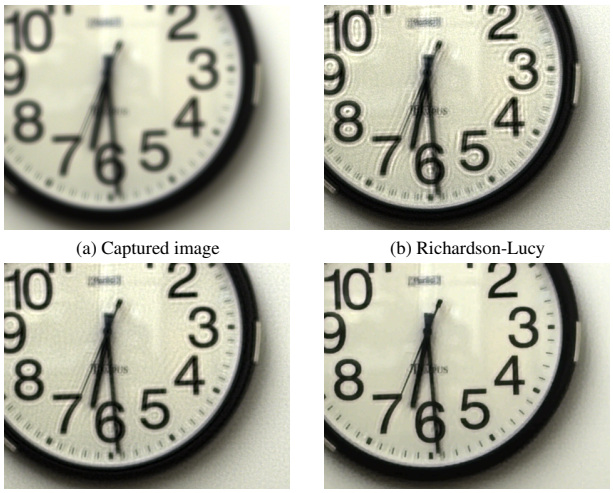
The probability model introduced in Section 2.1 allows us to detect the correct blur scale within an observed image window $y$. The correct scale should, in theory, be given by the model suggesting the most likely explanation: $k^* = argmax_k\ P_k(y)$. However, a variety of practical issues such as the high-frequency noise in the filter estimates mean that this proved to be unreliable. A more robust alternative is to use the unnormalized energy term $E_k(y) = y^T\Sigma_k^{-1}y$ from the model, in conjunction with a set of weightings for each scale: $k^* = argmin_k\ \lambda_k E_k(y)$. The weights $\lambda_k$ were learnt to minimize the scale misclassification error on a set of training images having a known depth profile. Since evaluating $y^T\Sigma^{-1}y$ is very slow, we approximate the energy term by the reconstruction error achieved by the ML solution:

$$y^T\Sigma_k^{-1}y \approx \frac{1}{\eta^2}|C_{f_k}x^*-y|^2 \quad (13)$$

where $x^*$ is the deblurred image, obtained by solving Eqn. 11.

## 4 Handling Depth Variations

If the captured image were filled by a planar object at a constant distance from the camera, the blur kernel would be uniform over



(a) Captured image (b) Richardson-Lucy

(c) Gaussian prior (d) Sparsity prior

**Figure 7:** *Comparison of deblurring algorithms applied to an image captured using our coded aperture. Note the ringing artifacts in the Richardson-Lucy output. The sparsity prior output shows less noise than the other two approaches.*

the image. In this case, recovering the sharp image would involve the estimation of a single blur scale for the entire image. However, interesting real world scenes include depth variations and so a separate blur scale should be inferred for every image pixel. A practical compromise is to use small local windows, within which the depth is assumed to be constant. However, if the windows are small the depth classification may be unreliable, particularly when the window contains little texture. This issue is common to most passive illumination depth reconstruction algorithms.

We start by deblurring the entire image with each of the scaled kernels (according to Eqn. 10), providing K possible decoded images $x_1,..,x_K$. For each scale, the reconstruction error $e_k = y - f_k * x_k$ is computed. A decoded image $x_k$ will usually provide a smooth plausible reconstruction for parts of the image where $k$ is the true scale. The reconstruction in other areas, whose depths differ from $k$, will contain serious ringing artifacts since those areas cannot be plausibly explained by the $k^{\text{th}}$ scale (see Figures 11 & 12 for examples of such artifacts). These artifacts ensure that the reconstruction error for such areas will be high. Using Eqn. 13 we compute a local approximation for the energy $E_k(y(i))$ around the $i$th image pixel, by averaging the reconstruction error over a small local window:

$$\hat{E}_k(y(i)) \approx \sum_{j \in W_i} e_k(j)^2 \tag{14}$$

The local energy estimate is then used to locally select the depth $d(i)$ in the $i$th pixel

$$d(i) = argmin_k \ \lambda_k \hat{E}_k(y(i)) \tag{15}$$

A local depth map is shown in Figure 8(b). While this local approach captures a surprising amount of information, it is quite noisy, especially for uniform texture-less regions. In order to produce a visually plausible deconvolved image, the local depth map is often sufficient, since the texture-less regions will not produce ringing when deconvolved with the wrong scale of filter. Hence we can produce a high quality sharp image by picking each pixel independently from the layer with smallest reconstruction error. That is, we construct the deblurred image as $x(i) = x_{d(i)}(i)$, using the local depth estimates $d(i)$ defined in Eqn. 15. Examples of deblurred images are shown in Figure 10.

However, to produce a depth estimate which could be useful for tasks like object extraction and scene re-rendering, the depth map has to be smoothed. We seek a regularized depth labeling $\bar{d}$ which will be close to the local estimate in Eqn. 15, but will also be smooth. Additionally, we prefer the depth discontinuities to align with the image edges. We formulate this as an energy minimization, using a Markov random field over the image, in the manner of classic stereo and image segmentation approaches (e.g. [Boykov et al. 2001])

$$E(\bar{d}) = \sum_i E_1(\bar{d}_i) + \nu \sum_{i,j} E_2(\bar{d}_i, \bar{d}_j) \tag{16}$$

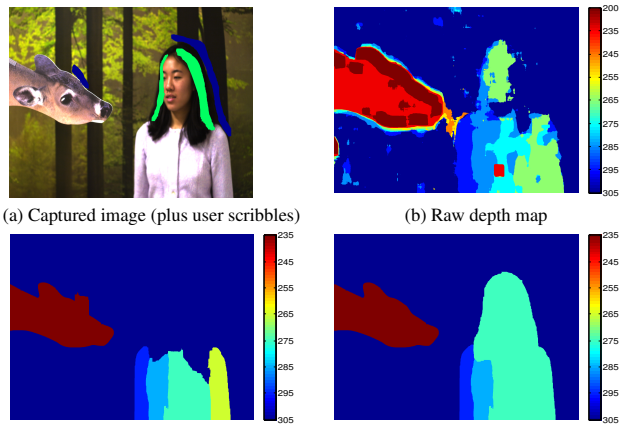
where the local energy term is set to

$$E_1(\bar{d}_i) = \begin{cases} 0 & \bar{d}_i = d_i \\ 1 & \bar{d}_i \neq d_i \end{cases}$$

There is also a pairwise energy term between neighboring pixels making depth discontinuities cheaper when they align with the image edges:

$$E_2(\bar{d}_i, \bar{d}_j) = \begin{cases} 0 & \bar{d}_i = \bar{d}_j \\ e^{-(y_i - y_j)^2/\sigma^2} & \bar{d}_i \neq \bar{d}_j \end{cases}$$

We then search for the minimal energy labeling as a min-cut in a graph. The resulting smoothed depth map is presented in Figure 8(c). Occasionally, the depth labeling misses the exact layer boundaries due to insufficient image contrast. To correct this, a user can apply brush strokes to the image with the required depth assignment. The strokes are treated as hard constraints in the Markov random field and result in an improved depth map, as illustrated in Figure 8(d).




(a) Captured image (plus user scribbles) (b) Raw depth map

(c) Graph cuts (d) After user correction

**Figure 8:** *Regularizing depth estimation*

## 5 Results

We first detail the physical construction and calibration of our chosen aperture pattern. Then we show a variety of real scenes, recovering both the depth map and fully sharp image. As a baseline experiment, we then compare the performances of conventional and coded apertures, using the same deblurring and depth estimation algorithms. Finally, we show some applications made possible by the additional depth information for each image, such as refocusing and scene re-rendering.

### 5.1 Calibration

The best performing filter under the criterion of Eqn. 8 was cut from gray card and inserted into an off-the-shelf Canon 50mm $f/1.8$ lens (shown in Figure3(b)) mounted on a Canon 20D DSLR. To calibrate

the lens the focus was locked at $D = 2$m and the camera was moved back until $D_k = 3$m in 10cm increments. At each interval, a planar pattern of random curves was captured. After aligning the focused calibration image with each of the blurry versions the blur kernel was deduced in a least-squares fashion, using a small amount of regularization to constrain the high-frequencies within the kernel. When $D_k$ is close to $D$ the blur is very small ($< 4$ pixels) making depth discrimination impossible due to lack of structure in the blur, although the image remains relatively sharp. For our setup, this "dead-zone" extends up to 35cm from the focal plane.

Since the lens does not perfectly obey the thin lens model, the kernel varies slightly across the image, the distortion being more pronounced in the horizontal plane. Consequently, kernels were inferred at 7 different horizontal locations within the image. The computed kernels at a number of depths are shown in Figure 9. To enable a direct comparison between a conventional and coded apertures, we also calibrated an unmodified Canon 50mm $f/1.8$ lens in the same fashion.

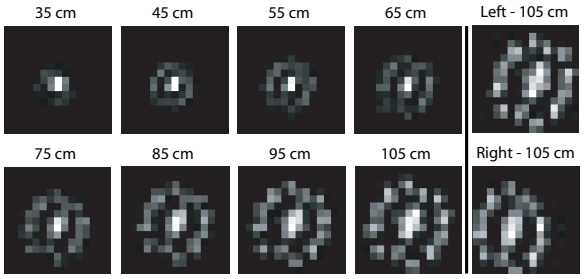



**Figure 9:** *Left: Calibrated kernels at a variety of depths from the focus plane. All are taken from the center of the frame. Right: Kernels from the far left and right of the frame at 1.05m from the focal place, showing significant radial distortion.*

## 5.2 Test Scenes

To evaluate our system we capture a number of 2 megapixel images of scenes whose depth varies over the same range used in calibration (between 2 and 3.05m from the camera). All the recovered images, unless otherwise indicated, utilized a sparse prior in deblurring. Owing to the high resolution of many of the results, we include full-sized versions in the supplementary material on the project webpage.

The table scene shown in Figure 1 contains objects spread at a variety of depths. The close-ups show the successful removal of the coded blur from the bottles on the right side of scene. The depth map (obtained without user assistance) gives a fairly accurate reconstruction of distance from the camera. For example, the central two beer bottles are placed only $5-10$ cm in front of the peripheral two, yet depth map still captures this difference.

Figure 10 shows two women sitting on a sofa. The depth map (produced without manual stroke hints) reveals that one is sitting back while the other is sitting forward. The tilted pose of the woman on the right results in the depth map splitting across her body. The depth errors in the background on the left are due to specularities which, aside from being saturated, originate from a different distance to the rest of the scene. The arms of the woman on the left have been merged into the background due to lack of distinctive high-frequency texture on them.

Note that the recovery of the all-focus image directly uses the local depth maps (as in Figure 8(b)) without regularization and without user corrections. Any ambiguities in the local depth map mean that that more than one blur scale gives a ringing free explanation (this is especially true for uniform image areas). Hence such errors in depth estimation will not result in visual artifacts. However, regularized depth maps were used for refocusing and novel view synthesis.

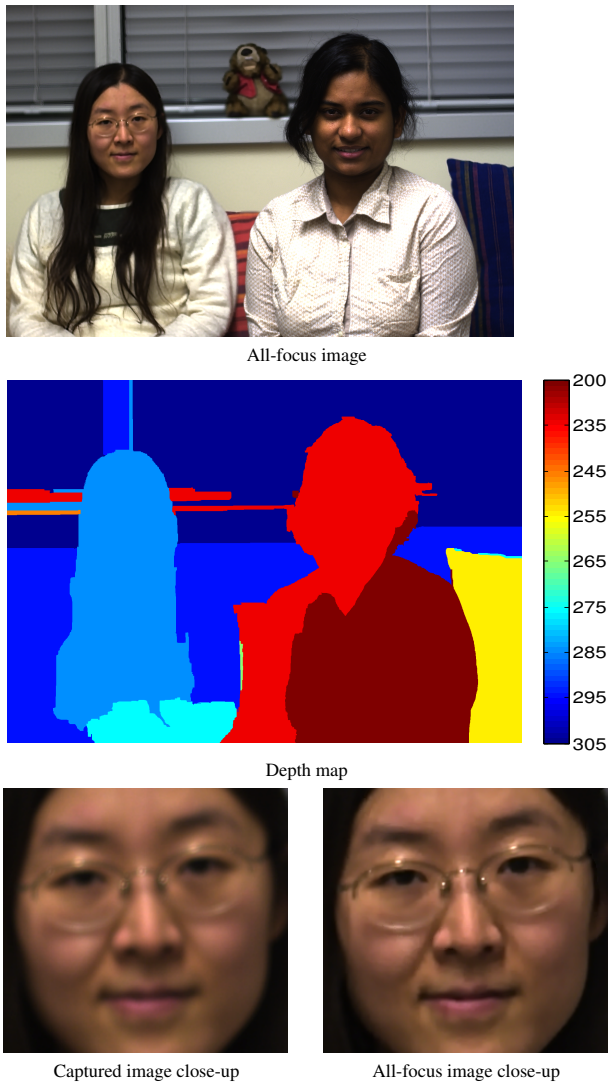



**Figure 10:** *The recovered sharp image of a sofa scene with two women and associated depth map. The close-up images show the extended depth of focus offered by our method.*

## 5.3 Comparison with a Conventional Aperture

To assess the importance of the coded aperture in our system, in Figure 12 we make a practical comparison between coded and conventional apertures. The same scene was captured with conventional and coded lenses and an all-focus image recovered using the appropriate set of filters obtained in calibration. The coded aperture result is mostly sharp whereas the conventional lens result shows significant artifacts in the foreground where the depth estimate is drastically wrong.

We also performed a quantitative comparison between the two aperture types using images of planar scenes of known depth, giving a evaluation of the robustness of our entire system. When considering local evidence alone, the coded aperture accurately classified the depth in 80% of the images while the conventional aperture accurately classified the depth only 40% of the time. These results

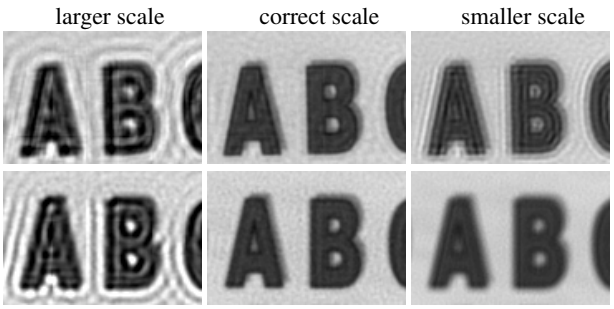



**Figure 11:** *Deblurring with varying blur scale. Top: coded aperture, Bottom: conventional aperture.*

validate the theoretical prediction from Figure 5 and justifies the use of a coded aperture over an unmodified lens.

To further illustrate the difference, Figure 11 presents image windows captured using conventional and coded lenses. Those windows were deblurred with the correct blur scale, too large a scale and too small a scale. With a coded lens shifting the scale in both directions generates ringing, however with a conventional kernel ringing occurs only in one direction. It should be noted that ringing indicates that the observed image can not be well explained by the proposed kernel. Thus with a conventional lens a smaller scale is also a legal explanation, leaving a larger uncertainty on the depth estimation. A coded lens, on the other hand, is better in nailing down the correct scale.

### 5.4 Applications

In Figure 13 we show how an all-focus image can be synthetically refocused to selectively pick out any of the individuals, in the style of Ng et al [2005]. The depth information can also be used to translate the camera location post-capture in a realistic manner, shifting each depth plane according to its distance from the camera. The new parts of the scene revealed by the motion are in-painted from neighboring regions using Photoshop's "Healing Brush" tool. A video demonstrating viewpoint translation as well as additional refocusing results can be found in the supplementary file and on the project webpage (`http://groups.csail.mit.edu/graphics/CodedAperture`).

## 6 Discussion

In this work we have shown how a simple modification of a conventional lens – the insertion of a patterned disc of cardboard into the aperture – permits the recovery of both an all-focus image and depth from a single image. The pattern produces a characteristic distribution of image frequencies that is very sensitive to the exact scale of defocus blur.

Like most classical stereo vision algorithms, the approach relies on the presence of a sufficient amount of texture in the scene. Robust segmentation of depth layers requires distinctive color boundaries between occlusion edges. In the absence of those, user assistance may be required.

While the ability to refocus post-exposure may lessen the need to vary the aperture size to control the depth of field, different aperture areas could be obtained using a fixed set of different aperture patterns. The insertion of the filter into the lens also reduces the amount of light that reaches the sensor. For the filter used in our experiments, around 50% of the light is blocked (i.e. one stop of exposure). We argue that this is an acceptable loss, given the extra depth information that is obtainable.


**Figure 12:** *Showing the need for a coded aperture. Recovered images using our coded aperture and the result of the same calibration and processing steps applied to a conventional aperture image. The unreliable depth estimates of the conventional aperture image lead to ringing artifacts in the deblurred image.*

Our approach requires an exact calibration of the blur filter over depth values. Currently, we have only calibrated our filter for a fixed focus setting over a relatively narrow range of depth values ($2-3$m from the camera). At extreme defocus values, the blur cannot be robustly inverted. A more general implementation will require calibration over a range of focus settings, and storing the focus setting with each exposure (a capability of many existing cameras).

## Acknowledgements

We are indebted to Ted Adelson for insights and suggestions and for the usage of his lab space. Funding for the project was provided by NGA NEGI-1582-04-0004 and Shell Research. Frédo Durand acknowledges a Microsoft Research New Faculty Fellowship and a Sloan fellowship.

## References

ADELSON, E. H., AND WANG, J. Y. A. 1992. Single lens stereo with a plenoptic camera. *IEEE Trans. Pattern Anal. Mach. Intell. 14*, 2, 99–106.

AXELSSON, P. 1999. Processing of laser scanner data–algorithms and applications. *ISPRS Journal of Photogrammetry and Remote Sensing 54*, 138–147.

BARRETT, R., BERRY, M., CHAN, T. F., DEMMEL, J., DONATO, J., DONGARRA, J., EIJKHOUT, V., POZO, R., ROMINE, C., AND DER VORST, H. V. 1994. *Templates for the Solution of*

**Figure 13:** *Refocusing: Using the recovered depth map and all-focus image, the user can refocus, post-exposure, to selected depth-layers.*

*Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*. SIAM, Philadelphia, PA.

BOYKOV, Y., VEKSLER, O., AND ZABIH, R. 2001. Fast approximate energy minimization via graph cuts. *PAMI 23* (Nov), 1222–1239.

CATHEY, W., AND DOWSKI, R. 1995. A new paradigm for imaging systems. *Applied Optics 41*, 1859–1866.

CHAUDHURI, S., AND RAJAGOPALAN, A. 1999. *Depth from defocus: A real aperture imaging approach*. Springer-Verlag, New York.

DOWSKI, E. R., AND CATHEY, W. T. 1994. Single-lens single-image incoherent passive-ranging systems. *Applied Optics 33*, 6762–6773.

FARID, H., AND SIMONCELLI, E. P. 1998. Range estimation by optical differentiation. *Journal of the Optical Society of America 15*, 1777–1786.

FAVARO, P., MENNUCCI, A., AND SOATTO, S. 2003. Observing shape from defocused images. *Int. J. Comput. Vision 52*, 1, 25–43.

FENIMORE, E., AND CANNON, T. 1978. Coded aperture imaging with uniformly redundant rays. *Applied Optics 17*, 337–347.

FERGUS, R., SINGH, B., HERTZMANN, A., ROWEIS, S. T., AND FREEMAN, W. 2006. Removing camera shake from a single photograph. *ACM Transactions on Graphics, SIGGRAPH 2006 Conference Proceedings, Boston, MA 25*, 787–794.

GEORGIEV, T., ZHENG, K. C., CURLESS, B., SALESIN, D., NAYAR, S., AND INTWALA, C. 2006. Spatio-angular resolution tradeoffs in integral photography. In *Rendering Techniques 2006: 17th Eurographics Workshop on Rendering*, 263–272.

GREENGARD, A., SCHECHNER, Y., AND PIESTUN, R. 2006. Depth from diffracted rotation. *Optics Letters 31*, 181–183.

GROSSMANN, P. 1987. Depth from focus. *Pattern Recognition Letters 5*, 1 (Jan.), 63–69.

HASINOFF, S. W., AND KUTULAKOS, K. N. 2006. Confocal stereo. In *European Conference on Computer Vision*, I: 620–634.

HIURA, S., AND MATSUYAMA, T. 1998. Depth measurement by the multi-focus camera. In *CVPR*, IEEE Computer Society, 953–961.

JONES, D., AND LAMB, D., 1993. Analyzing the visual echo: passive 3-D imaging with a multiple aperture camera. Technical Report CIM 93-3, Dept. of Electrical Engineering, McGill University.

KUNDUR, D., AND HATZINAKOS, D. 1996. Blind image deconvolution. *IEEE Signal Processing Magazine 13*, 3 (May), 43–64.

LAI, S.-H., FU, C.-W., AND CHANG, S. 1992. A generalized depth estimation algorithm with a single image. *IEEE Trans. Pattern Anal. Mach. Intell. 14*, 4, 405–411.

LEVIN, A., AND WEISS, Y. To appear. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

LEVIN, A. 2006. Blind motion deblurring using image statistics. In *Advances in Neural Information Processing Systems (NIPS)*.

LEVOY, M., NG, R., ADAMS, A., FOOTER, M., AND HOROWITZ, M. 2006. Light field microscopy. *ACM Transactions on Graphics 25*, 3 (July), 924–934.

NAYAR, S. K., WATANABE, M., AND NOGUCHI, M. 1995. Real-time focus range sensor. In *ICCV*, 995–1001.

NG, R., LEVOY, M., BREDIF, M., DUVAL, G., HOROWITZ, M., AND HANRAHAN, P. 2005. Light field photography with a hand-held plenoptic camera. *Stanford University Computer Science Tech Report CSTR 2005-02*.

OLSHAUSEN, B. A., AND FIELD, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature 381* (June), 607–609.

PENTLAND, A. P. 1987. A new sense for depth of field. *IEEE Trans. Pattern Anal. Mach. Intell. 9*, 4, 523–531.

PREMARATNE, P., AND KO, C. C. 1999. Zero sheet separation of blurred images with symmetrical point spread functions. *Signals, Systems, and Computers*, 1297–1299.

RASKAR, R., AGRAWAL, A., AND TUBMLIN, J. 2006. Coded exposure photography: Motion deblurring using fluttered shutter. *ACM Transactions on Graphics, SIGGRAPH 2006 Conference Proceedings, Boston, MA 25*, 795–804.

SCHARSTEIN, D., AND SZELISKI, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Intl. J. Computer Vision 47*, 1 (April), 7–42.

ZHANG, L., AND NAYAR, S. K. 2006. Projection defocus analysis for scene capture and image display. *ACM Trans. on Graphics (also Proc. of ACM SIGGRAPH)* (Jul).