Depth and Deblurring from a Spectrally-varying Depth-of-Field

Depth and Deblurring from a Spectrally-varying Depth-of-Field

Ayan Chakrabarti and Todd Zickler

Harvard University, Cambridge, MA, USA.
ayanc@eecs.harvard.edu, zickler@seas.harvard.edu

Abstract. We propose modifying the aperture of a conventional color camera so that the effective aperture size for one color channel is smaller than that for the other two. This produces an image where different color channels have different depths-of-field, and from this we can computationally recover scene depth, reconstruct an all-focus image and achieve synthetic re-focusing, all from a single shot. These capabilities are enabled by a spatio-spectral image model that encodes the statistical relationship between gradient profiles across color channels. This approach substantially improves depth accuracy over alternative single-shot coded-aperture designs, and since it avoids introducing additional spatial distortions and is light efficient, it allows high-quality deblurring and lower exposure times. We demonstrate these benefits with comparisons on synthetic data, as well as results on images captured with a prototype lens.

1 Introduction

Coded-aperture depth-from-defocus (DFD) techniques (e.g., [1, 2]) have significantly improved our ability to computationally recover depth from a single image. The recovered depth information can be used to deblur the observations and generate a sharp image of the scene, and this provides post-capture opportunities for extending depth of field, changing focus, and creating synthetic views—all from a single exposure. The key idea in these methods is to control the optical blur induced by defocus by inserting a coded pattern into the aperture of a conventional camera. The pattern is designed so that images of scene patches at different depths exhibit distinctive statistical spatial structure, and this improves one's ability to discriminate between depth levels.

Existing single-shot DFD techniques recover depth by relying on statistical models that encode the spatial structure of sharp natural images. This is necessary because the problem is ill-posed, with both the depth map and the scene radiance being unknown. Depth accuracy is therefore limited by the inherent variability of the visual world, and even though existing aperture codes dramatically improve depth discrimination, user intervention is still often required to produce a reliable depth map [1]. Furthermore, the improvement in depth discrimination comes at a cost: the quality of the sharp image that can be obtained from the depth map and the input observations is diminished because (1) a coded aperture transmits less light than a regular aperture; and (2) the blur

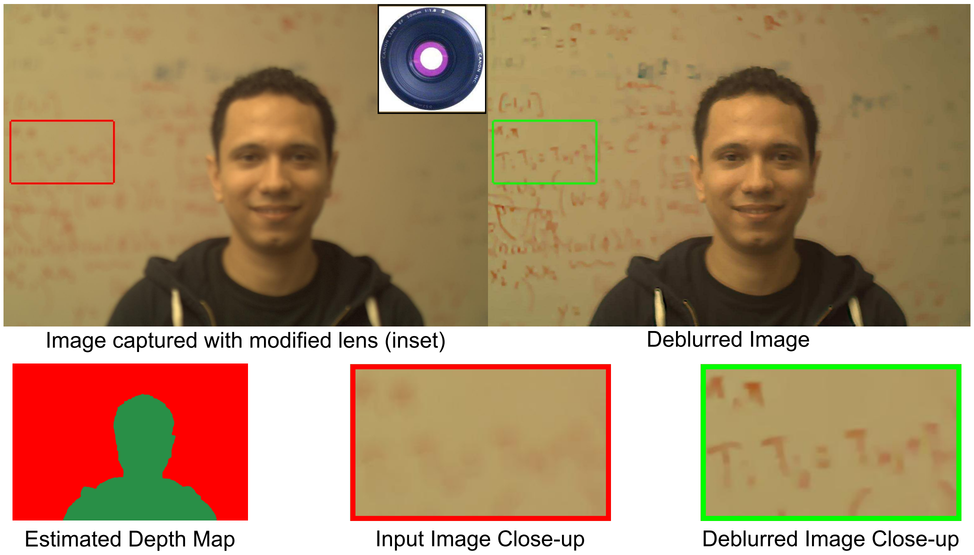


Image captured with modified lens (inset)

Deblurred Image

Estimated Depth Map

Input Image Close-up

Deblurred Image Close-up

Fig. 1. The proposed color-coded aperture generates reliable estimates of scene depth with greater accuracy than color-neutral approaches (no user interaction is required), while simultaneously allowing high-quality deblurring.

it induces is harder to invert [3]. These drawbacks have motivated consideration of acquisition systems that compromise on the convenience of a single exposure, and sequentially acquire multiple images with complimentary aperture codes [3].

This paper proposes an alternative aperture design and estimation approach for single-shot DFD based on the following premise: while spatial gradients in any individual color channel exhibit substantial variability, there is a stronger statistical relationship *between* spatial gradient profiles across different color channels. We exploit this relationship for depth recovery by inserting a ring-shaped color filter in the camera aperture (see Fig. 1), thereby inducing different depths-of-field in different color channels. The channels in the recorded image are affected by different degrees of defocus blur, and by appropriately defining a spatio-spectral image prior, we can reliably estimate depth from this spectrally-varying defocus. We show empirically that this approach yields depth estimates that are more accurate than those of existing single-shot DFD methods, and that it enables the automatic recovery of depth without user intervention. We also show that it can provide high-quality sharp images— both because the aperture pattern is more light efficient and because the approach relies on introducing spectral blur variation instead of increased spatial distortion— and that deblurring can be achieved efficiently through the use of a color-adapted version of fast deconvolution approaches based on half-quadratic splitting [4].

1.1 Related Work

Cameras with finite apertures record images that are affected by depth-dependent defocus blur, and so blur can be used as a cue for recovering depth. To obtain depth this way, traditional DFD approaches capture multiple images of the same scene with varying focal distance (e.g., [5]) or aperture size [6]. When the camera is calibrated, this is often sufficient to disambiguate between depth and surface radiance. In contrast, estimation with a single image is considerably more ill-posed. To address this, coded-aperture techniques [1, 2] use a prior model for the spatial structure of natural images—usually a probability distribution on spatial gradients of greyscale images— and an aperture mask that is designed from this prior. Together, the mask and the prior significantly improve single-shot depth estimation over what is possible with a regular un-coded aperture.

However, coded single-shot approaches still struggle to handle the variability in real-world images, and often require manual intervention to generate reliable depth maps [1]. Moreover, one's ability to use the observations and the recovered depth to subsequently recover a sharp *all-focus* image is diminished because the spatial distortion introduced by the code for depth discrimination has the undesirable side-effect of attenuating important spatial frequency content [3]. These limitations have inspired researchers to re-consider the less convenient multi-shot DFD, where distinct and complimentary aperture codes are applied in sequential exposures for reliable deblurring and depth estimation [3, 7].

A key observation is that these existing coded DFD techniques ignore color information by using color-neutral codes and leveraging spatial statistics alone.We show that a statistical model encoding joint *spatio-spectral* image structure is significantly more powerful for single-shot DFD. It enables depth estimation with greater accuracy than a color-neutral aperture optimized for the same task, while allowing the same quality of deconvolution as a regular un-coded aperture.

Recently, Cossairt and Nayar [8] proposed a color-based approach to invert defocus blur without explicitly estimating depth. They assert that in images acquired using a lens with significant chromatic aberration, the effective blur kernel for the luminance channel (i.e., the mean across color channels) can be treated as constant for depth values that lie within the range of per-wavelength focal distances for the lens. A sharper image can then be recovered by deconvolving the luminance channel with this kernel, although this leaves the chrominance channels with residual blur and chromatic aberration. In contrast, we seek to recover scene depth explicitly which allows us to deblur all color channels when forming the all-focus image, and is useful for applications such as synthetic refocusing.

Our approach relies on inserting a color filter pattern into the aperture plane, a property that it shares with some existing single-shot techniques for stereo-based depth estimation [9–11]. In these methods, the aperture is divided into non-overlapping per-channel apertures so that the observed color channels experience parallax. Depth is estimated by assuming that pixel colors of an aligned image will be correlated. An important limitation of this approach is that dividing a regular aperture into three non-overlapping ones blocks a large amount of light. For example, assuming ideal color filters, the aperture pattern of Bando et al. [11] transmits only 16% of the incident light, compared to 40% for the color-neutral pattern proposed in [1], and 78% for the design proposed in this paper. Moreover, our experiments indicate that even with longer exposure times to compensate for light attenuation, the stereo-based approach does not provide better depth accuracy than traditional DFD-based patterns (see Sec. 6).

Bando et al. [11] recover depth using a color image model defined in terms of statistics of local per-pixel color distributions. Similar models have been used for segmentation [12], matting [13], restoration [14], and so on. In contrast, we model properties of *gradients* across color channels, since these spatio-spectral statistics [15] are better suited for analyzing the effects of blur.

2 Problem Formulation and Camera Design

Based on the thin lens model, the projection $Y(n)$ of a fronto-parallel surface patch observed by a color camera at image location $n \in \mathbb{R}^2$, can be expressed as the convolution of a latent sharp image $X(n)$ of that patch with a depth-dependent blur kernel $k_{r(n)}$:

$$y^{\{i\}}(n) = (x^{\{i\}} * k_{r(n)})(n), \forall i \in \{R, G, B\}, \tag{1}$$

where k_r is a scaled version of the camera's aperture shape, and $y^{\{i\}}$ and $x^{\{i\}}$ are the i^{th} color channels of Y and X respectively. We assume in this paper that the blur kernels k_r are circular pill-box kernels of radius r.

The effective blur radii $r(n)$ in pixels can be related to scene depth $d(n)$, focal length f, and aperture radius A as $r(n) \propto A(d(n) - f)/f$, where the constant of proportionality depends on camera parameters. For negative values of r, the kernel k_r corresponds to a mirrored version of the aperture shape. But when the aperture is symmetric, $k_r = k_{-r}$, and this induces an ambiguity between depth values at equal distances in front of and behind the plane of focus. As in [1], we assume that during capture the camera is focused to the nearest surface or closer, i.e., $d(n) \geq f$. With a calibrated camera and known focal length, the blur radius $r(n)$ can therefore be directly related to the scene depth $d(n)$, and the problem of depth estimation reduces to one of recovering $r(n)$. While this model is based on the simple thin lens model, and neglects diffraction effects and lens distortion, we find it to be a reasonable approximation when imaging surfaces at a reasonable distance from the camera ($>$ 1m for a standard 50mm lens).

2.1 Aperture Modification

The aperture pattern we propose is designed to induce a different defocus blur kernel $k_r^{\{i\}}$ in each color channel i. As illustrated in Fig. 2, we construct this pattern by cutting a ring shape from a color filter that, in the ideal case, attenuates all light in one of the recorded color channels while perfectly transmitting the other two. In our design, we choose to attenuate the green channel which is typically recorded with the highest signal-to-noise ratio (SNR) by digital camera sensors. The outer boundary of the mask has radius equal to that of the lens aperture A, and the inner boundary has a smaller radius: αA, $\alpha < 1$. This inner radius acts as the effective aperture radius for the green channel, while the other channels remain unaffected and are imaged with the full aperture radius A. Images captured with this mask therefore have a spectrally-varying depth of field, with a larger depth of field in the green channel than in red and blue.

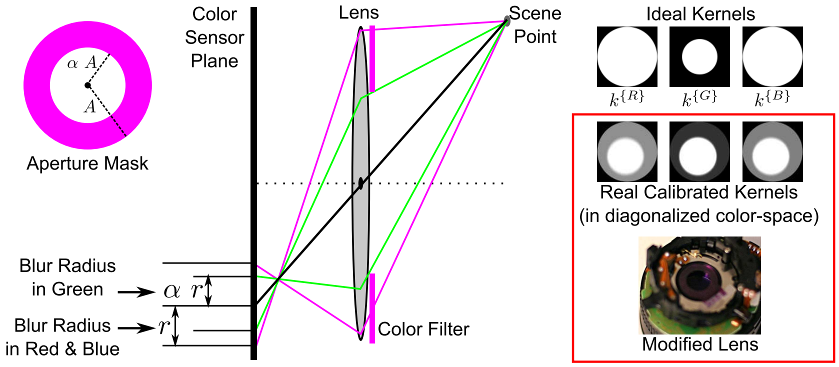


Fig. 2. Aperture Modification. A ring-shaped color mask is placed in the lens aperture to induce defocus blur that varies across color channels. Ideally, the filter attenuates only the green channel inducing a more compact defocus blur kernel in that channel for any depth. This is approximated using a readily-available color filter, which induces per-channel defocus blur kernels (at a specific depth) shown in the red box, right.

A low value for the parameter α causes greater spectral variation in the depth of field and provides a stronger depth cue, but also reduces light efficiency by a factor of $(1-\alpha^2)/3$ (assuming an ideal color filter with perfect attenuation in one channel and perfect transmission in the other two). Through experiments with a digital SLR camera, we find that $\alpha = 0.59$ provides a good balance between these concerns, and corresponds to an efficiency of 78%. Interestingly, this choice of parameter α is close to the optimal ratio of pill-box radii for two-shot DFD techniques [3, 16], even though our depth estimation approach is quite different.

2.2 Prototype Lens

To evaluate our design, we constructed a prototype by modifiying a Canon EF 50mm F/1.8 II lens. A mask was cut out of a sheet of the “Roscolux CalColor-60-Magenta” filter based on the specifications above, and placed in the aperture diaphragm of the lens (see Fig. 2). This filter was chosen because it is readily available, and has spectral characteristics that are reasonably close to our requirements, with a significantly higher attenuation in the green channel than in red and blue. However, the constructed mask has a reduced light-efficiency of about 60%. While this serves to demonstrate the efficacy of our approach, we recommend ultimately using camera sensors and a color aperture filter that are jointly constructed to better match the design criteria.

For calibration, we first estimate the spectral transmission of the color filter separately using images of a color checker chart taken with a regular lens, with and without the color filter in front of the camera. The effect of the filter is modeled as a diagonal linear transform M in a modified color space V: $V^T X' = M V^T X$, where X' and $X \in \mathbb{R}^3$ are camera sensor responses with and without the filter, and V is a unitary matrix. We use the transformed color space defined by V for both depth estimation and deblurring.

Next, we capture sharp and blurred image pairs of a calibration target with the modified lens to estimate the relative location and radius of the “inner ring” with respect to the full aperture. As shown in Fig. 2, the final calibrated kernels (in the transformed color space) are formed by modifying a regular pill-box kernel to account for per-channel attenuation outside this inner ring.

3 Image Model

To recover depth from images captured using the proposed color-coded aperture, we use a spatio-spectral model for images of real-world scenes. In particular, we define an image model that characterizes the relationship between local gradient profiles across color channels. An important distinction of our approach is that we do not rely on the sharpness of these gradient profiles, as is done in existing single-shot DFD approaches [1], and other blur estimation applications (e.g., [17]). Instead, we rely only on the *agreement* of the gradient profiles across color channels, whether or not they are sharp. Therefore, the proposed model describes any real-world image in which all channels are affected by the same blur, while being sensitive to spectral variations in blur.

Let $X^{\nabla}(n) \in \mathbb{R}^3$ be the *color gradient* vector at pixel n, obtained by applying an oriented gradient filter ∇ to each color channel:

$$X^{\nabla}(n) = \left[(x^{\{1\}} * \nabla)(n), (x^{\{2\}} * \nabla)(n), (x^{\{3\}} * \nabla)(n)\right]^T. \tag{2}$$

We consider a spatial profile of these gradient vectors $\{X^{\nabla}(n)\}_{n\in\mathcal{W}}$, over a local one-dimensional (1D) window $\mathcal{W}$ that has the same orientation as the gradient filter, and we create our model based on the expectation that spatial profiles of the different color gradient channels $(x^{\{i\}} * \nabla)(n)$ will be scaled versions of one another. This can be represented using a generative probabilistic model that factorizes these vectors into spatial and spectral components as

$$X^{\nabla}(n) = S\ t(n) + Z(n), \qquad Z(n) \sim \mathcal{N}(0, \sigma_z^2 I_{3\times 3}), \tag{3}$$

where $S \in \mathbb{R}^3$ contains the latent per-channel scale factors; $t(n) \in \mathbb{R}$ is the latent common spatial profile; and $Z(n)$ is white Gaussian noise with variance σ_z^2. This is equivalent to expecting that the different color gradient vectors in the window $\mathcal{W}$ will lie on a line through the origin in $\mathbb{R}^3$.

Figure 3 motivates this model using two example windows from a typical color image. When the window $\mathcal{W}$ back-projects to a region containing the same diffuse material (Fig. 3 (right)), color gradients are induced by scalar changes in shading that do not affect chromaticity. In this case, the model fits with S proportional to the diffuse material color. On the other hand, when $\mathcal{W}$ spans a material boundary (Fig. 3 (left)), the common spatial profile $t(n)$ encodes the shape of the edge between the two materials (which are typically large relative to the shading gradients within the materials on either side), and S encodes the contrast between their colors. Note that these properties also hold when the image is blurred with a spectrally-uniform kernel since this only affects the profile $t(n)$ (Fig. 3 (bottom)). Now, if one channel— green in our design— is

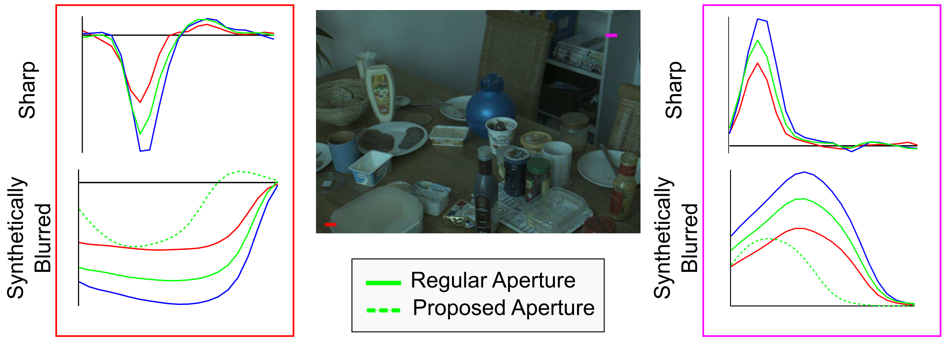


Fig. 3. Examples of gradient profiles in a linear RGB image (source: [18]), corresponding to a material boundary (left, red), and a homogeneous region (right, magenta). The profiles in different channels are seen to be approximately scaled versions of each other, in both the sharp image (top), and in the presence of spectrally-uniform blur (bottom). However, when the green channel image is blurred with a comparatively smaller blur (as is the case with the proposed aperture in Fig. 2), the corresponding profile, shown with a dotted line, no longer matches the profiles in the other two channels.

blurred less than the other two, the corresponding gradient profile (dotted green in Fig. 3) has a distinctly different shape, and the gradient vectors $X^{\nabla}(n)$ can not be factorized as in (3). The intuition above ignores many effects present in natural images— specular highlights, lighting variation, high-frequency texture, etc. However, this simple model proves to be adequately sensitive to the presence of spectral variation in blur, and enables reliable depth estimation.

Next, we define the function $\mathcal{L}_{\nabla}$ to measure the likelihood of an observed image window $\mathcal{W}$ under the model in (3) as

$$\mathcal{L}_{\nabla}(X, \mathcal{W}) = \max_{S, \{t(n)\}} \log p\left(\{X^{\nabla}(n)\} \,|S, \{t(n)\}\right). \tag{4}$$

The maximization above essentially corresponds to fitting a line to the gradient vectors $\{X^{\nabla}(n)\}$, and therefore (4) simplifies (up to a constant) to

$$\mathcal{L}_{\nabla}(X, \mathcal{W}) = -\left(\mathrm{tr}(\varLambda) - \lambda_1(\varLambda)\right), \quad \varLambda = \sum_{n \in \mathcal{W}} X^{\nabla}(n) X^{\nabla}(n)^T, \tag{5}$$

where $\mathrm{tr}(\varLambda)$ is the trace of $\varLambda$, and $\lambda_1(\varLambda)$ its largest eigenvalue. Note that a complete eigen-decomposition of $\varLambda$, which would be expensive if required for every patch, is not needed to compute (5). The largest eigen-value $\lambda_1(\varLambda)$ can be obtained using power iterations, and since $\varLambda$ is a 3×3 matrix, the method typically converges with sufficient accuracy in just three or four iterations.

Finally, we pool likelihoods from gradient profiles over multiple orientations at each location:

$$L(X, n) = \sum_{\theta} \mathcal{L}_{\nabla_\theta}(X, \mathcal{W}_\theta(n)), \tag{6}$$

where ∇_θ refers to a gradient filter at orientation θ, and $\mathcal{W}_\theta(n)$ to a window at the same orientation centered at location n. In our implementation, we use Gaussian-derivative filters (with a standard-deviation of one) at four orientations at multiples of 45°, and a window size of fifteen pixels.

4 Depth Estimation

In this section, we describe a method to recover scene depth from an image $Y(n)$ captured with the proposed color-coded aperture. Our goal is to estimate a depth value $d(n) \in \mathcal{D}$ at each image location n, where $\mathcal{D}$ is a discrete set of candidate depth levels. In our current implementation, we consider a set of thirteen levels corresponding to aperture radii $\{r_d : d \in \mathcal{D}\}$ ranging from 0 (i.e., no blur) to 24 pixels in steps of two pixels (depth accuracy is limited by estimation ability beyond this level of quantization). We employ a two-step estimation approach: we first generate local depth estimates using the color model in Sec. 3, and then combine these estimates with a smoothness constraint in a Markov random field (MRF) framework to segment the image into layers of constant depth.

4.1 Local Inference

Local depth estimates are computed by evaluating how well gradient profiles in each region of the observed image are explained by spectral variation in the induced defocus blur at each hypothetical depth level. We generate candidate blur-*aligned* versions $\tilde{Y}_d(n)$ of the observed image $Y(n)$ for each $d \in \mathcal{D}$, by compensating for the variation between the different per-channel kernels $\{k_{r_d}^{\{i\}}\}$, and then evaluate the likelihood of $\tilde{Y}_d(n)$ under our model. Without loss of generality, let the blur kernel $k_{r_d}^{\{R\}}$ for the red channel be equal or *larger* than the kernels for the other two. To construct $\tilde{Y}_d(n)$, we assume that each channel $y^{\{i\}}(n)$ in the observed image was blurred with the kernel that is known to occur in its channel at depth d, $k_{r_d}^{\{i\}}$, and compensate for the lower degrees of defocus in the blue and green channels. Intuitively, this can be achieved by deconvolving $y^{\{i\}}(n), i \in \{G, B\}$ with $k_{r_d}^{\{i\}}$, and then convolving both channels with $k_{r_d}^{\{R\}}$. With an ideal aperture pattern, only the green channel would require correction.

Since the overall effect of the correction $k_{r_d}^{\{R\}} * (k_{r_d}^{\{i\}})^{-1}$ is to increase the amount of defocus in the green and blue channels, we find that it can be reliably achieved by single convolutions with filters having limited support, ensuring that the resulting depth estimates $\hat{d}(n)$ will be well-localized. These corrective kernels, denoted $k_{r_d}^{\{R/i\}}, i \in \{G, B\}$, are constructed in the Fourier domain:

$$k_{r_d}^{\{R/i\}} = \mathcal{F}^{-1}\left[\frac{\mathcal{F}\left[k_{r_d}^{\{R\}}\right]\mathcal{F}\left[k_{r_d}^{\{i\}}\right]^*}{\left|\mathcal{F}\left[k_{r_d}^{\{i\}}\right]\right|^2 + \lambda_k\left(|\mathcal{F}[d_x]|^2 + |\mathcal{F}[d_y]|^2\right)}\right], \tag{7}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the $K \times K$ forward and inverse discrete Fourier transforms, with K being the desired spatial extent (in pixels) for the corrective kernels. The second term in the denominator of (7) serves to regularize the deconvolution step. Here, d_x and d_y denote horizontal and vertical finite-difference filters (i.e., $[-1, 1]$), and λ_k is the regularization weight (10^{-3} in our implementation).

The corrective kernel size K in (7) must be large enough to prevent aliasing in the Fourier domain, and depends on variation in the Fourier spectra of the different per-channel kernels $k_r^{\{i\}}$. After inspecting the corrective kernels generated with different choices, we choose K to be five times the support of the largest channel kernel $k^R_{r_d}$ for the ideal design, and half that size for kernels induced by the prototype lens. (The latter exhibit less variation between channels because of the non-ideal color filter and can therefore be corrected with a smaller corrective kernel, but require correcting two channels instead of one.) Note that the per-depth corrective kernels need only be computed once after lens calibration.

Once the aligned images $\tilde{Y}_d(n)$ have been constructed for each depth level, the local depth estimates $\hat{d}(n)$ are computed as:

$$\hat{d}(n) = \arg\max_{d \in \mathcal{D}} L(\tilde{Y}_d, n), \tag{8}$$

where L is the likelihood measure defined in (6). In addition to choosing the most likely depth, we also seek to characterize the confidence of this choice. For example, in overly smooth regions, the observed gradient profiles can be dominated by image noise and lead to arbitrary local depth assignments. To detect this, we construct an image $\tilde{Y}_\phi$ by simply blurring the green and blue channels of the observed image with a low-pass filter (we use a circular pill-box filter with support K in our implementation). This does not correspond to correction for any physically plausible depth value, and therefore serves as the "null hypothesis". In the next section, we will gauge the reliability of the local depth estimates $\hat{d}(n)$ from (8) by comparing its likelihood, $\hat{L}(n) = L(\tilde{Y}_{\hat{d}(n)}, n)$, to that of the null hypothesis, $L_\phi(n) = L(\tilde{Y}_\phi, n)$.

4.2 Regularized Depth Map Estimation

We generate a layered depth map $\bar{d}(n)$ by combining the local estimates above with a smoothness cost in an MRF model. This model is defined on a four-connected grid over the image, and the depth map $\bar{d}(n)$ is computed to minimize an energy function defined as

$$E(d(n)) = \sum U(d(n), n) + \eta \sum V_{n,n'}\left(d(n), d(n')\right), \tag{9}$$

where η is the relative weight of the smoothness term V to the local term U.

We use a clipped linear cost for the local energy term $U(d(n), n)$:

$$U(d, n) = \min\left(\left| r_{\hat{d}(n)} - r_d \right|, 8\right), \text{ if } \hat{L}(n) > \kappa\ L_\phi(n), \tag{10}$$

and zero otherwise (indicating a lack of confidence in the local estimate). We prefer this smooth cost for U over a zero-one indicator (i.e., $\delta[\hat{d} \neq d]$) as used in [1], because it makes the minimization of E less prone to local-optima. This is supported by our observation that even when the local estimates $\hat{d}(n)$ are not perfectly accurate, they are close to the true depth value (see Fig. 4).

Like [1], we use a smoothness cost V that encourages depth layer boundaries to align with image edges:

$$V_{n,n'}(d_1, d_2) = \begin{cases} 0, & \text{if } \ d_1 = d_2, \\ \exp(-\|\nabla_n\|^2/\sigma^2), & \text{otherwise,} \end{cases} \tag{11}$$

where $\|\nabla_n\|^2$ is the gradient energy at n, and the parameter σ is set to four times the median of $\|\nabla_n\|$ values across the image.

We solve for the depth map $\bar{d}(n)$ using graph-cuts, with the α-expansion step to handle multiple labels [19]. However, since the energy in (9) is not convex, this computation is expensive and often requires multiple restarts when the number of candidate labels is large. Therefore, we solve for $\bar{d}(n)$ over a reduced set of depth values that correspond to peaks of the histogram of the local depth estimates $\hat{d}(n)$, over the entire image. After computing a solution over this reduced set, the depth value for each layer (i.e., each connected region with constant depth) is set to the most frequent value of local depth $\hat{d}(n)$ in that layer. As a final post-processing step, we reassign any isolated regions with an area smaller than 1% of the image size to an adjoining layer with the closest depth value.

5 Deblurring and Refocusing

We can now generate a sharp image of the captured scene, where all regions appear to be in focus, by deconvolving each layer of the image with the per-channel kernels corresponding to its estimated depth. To deal with depth discontinuities, regions near a boundary between different depth layers are assigned the smallest of these depths. This simple approach proves effective because unlike the kernels in [1], those induced by our aperture are such that deconvolution assuming a *smaller* blur radius does not cause ringing artifacts, although some regions near layer boundaries do remain smooth in the deblurred image.

For each depth layer in the image with depth d, we first estimate $X_d(n)$ by deconvolving $Y_d(n)$ with the corresponding per-channel kernels $\{k_{r_d}^{\{i\}}\}_i$. Here, $Y_d(n)$ is formed from the observed image $Y(n)$ by blurring regions outside that layer, i.e., $Y_d(n) = Y(n)$ if $\bar{d}(n) = d$, and $y_d^{\{i\}}(n) = (y^{\{i\}} * k_{r_d}^{\{i\}})(n)$ otherwise. This ensures that the deblurred layer in $X_d(n)$ is not affected by ringing artifacts from neighboring regions. For deconvolution, we employ a new sparse regularization-based algorithm described in a supplementary technical report [20]. This method extends an existing fast deconvolution algorithm [4] to enforce local consistency in gradient colors. The expected chromaticity of each gradient is computed from an initial over-smoothed estimate of the sharp image (that is inexpensive to compute), and used as a hard constraint in each iteration of [4].

The different $X_d(n)$ are combined to generate a sharp image of the entire scene as $X(n) = X_{\bar{d}(n)}(n)$. This sharp image can subsequently be used to create synthetic views, where each depth layer is assigned an arbitrary degree of defocus. These generated views can correspond to different layers being in focus, and with a user-specified depth of field. We use spectrally-uniform circular pill-box kernels to blur each layer since this simulates capture with a regular aperture.

6 Experimental Results

We begin our evaluation by comparing the depth estimation accuracy of our approach to the color-neutral DFD design of [1] and the color-based single-

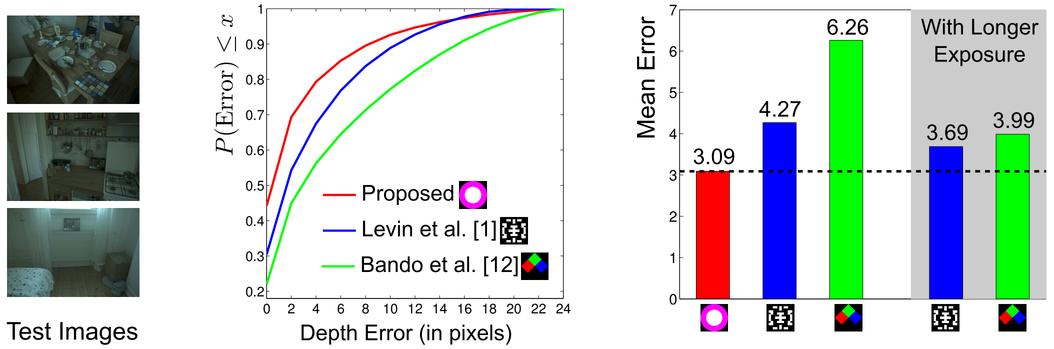


Fig. 4. Depth estimation performance for different single-shot coded-aperture methods, using synthetic experiments on three color images (left). Estimation error is measured as the absolute difference between true and estimated aperture radii in pixels. We show the error distribution for different approaches (center), as well as the mean error (right). For the latter, we also show performance with higher exposure times for the other designs, to compensate for the lower light efficiency of their aperture patterns.

shot stereo approach of [11]. Figure 4 shows results of synthetic evaluation on three linear color images from a public database [18]. We create synthetically blurred versions of these images at all candidate depth levels for each technique, through convolution with the kernels corresponding to their (ideal) aperture patterns followed by addition of Gaussian noise. We then use the respective *local* estimation algorithms (without MRF-based regularization) to estimate depth at each location. Bando et al. [11] recommend a window size of 15×15, and we use the same size for [1], as well as 1D windows of length 15 for our approach. Note that we use the same three test images for the training step (for parameters "λ_k") in [1], which may give an optimistic estimate of their performance.

We measure depth estimation error using the absolute difference between the aperture radii in pixels at the true and estimated depth values. We show the distribution and mean value of this error for each method in Fig. 4 (over roughly 10^8 samples over the three images and all simulated depth levels). The proposed approach offers a distinct advantage in accuracy over the other techniques. We also show mean error values for [1] and [11] by simulating higher exposure times to compensate for the lower light efficiency to generate images with SNR equivalent to our aperture. While this leads to an improvement in accuracy, the resulting mean errors are still higher than that from our design.

Next, we compare deblurring performance for images captured with our aperture to those captured with a regular aperture and the coded aperture in [1]. We show deconvolution results on synthetically blurred versions of a natural image that simulate capture for these apertures, assuming a known uniform depth, and equal exposure times for all three. Figure 5 shows the synthetically blurred images and deconvolution results, along with reconstruction quality (in terms of PSNR values) for the latter. We find that our design closely matches the performance with a regular aperture, with visually indistinguishable results and a negligible drop in PSNR. In contrast, the image recovered from [1] exhibits a comparatively higher loss of detail. In summary, the proposed design recovers

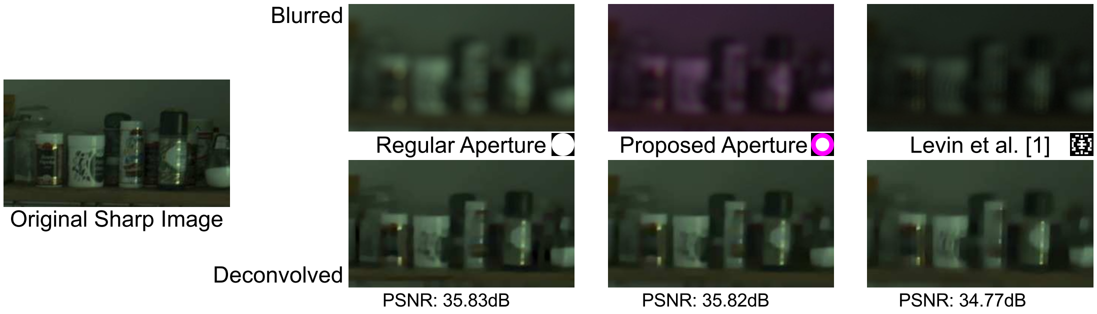


Fig. 5. Deconvolution performance. We compare deconvolution ability for images captured with different apertures, when the true depths are known. Deblurred estimates from the proposed aperture are almost identical to those from a regular aperture.

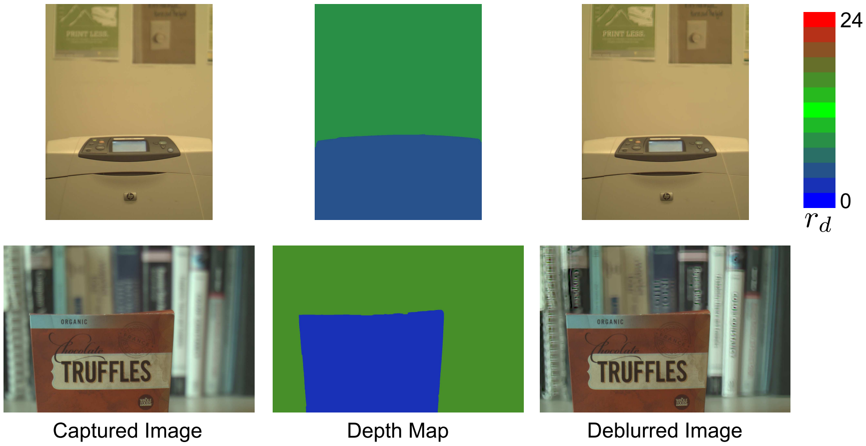
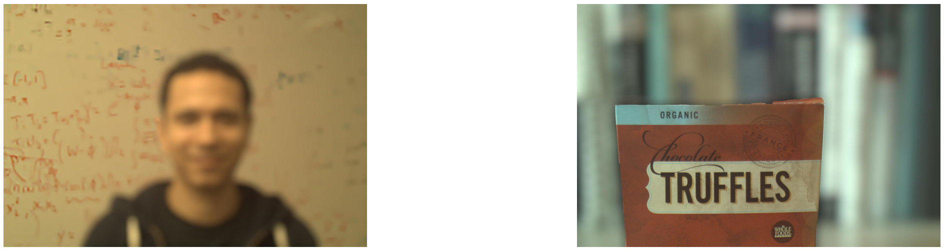


Fig. 6. Estimated depth maps and all-focus images, for two additional scenes captured with the prototype lens. We recover reliable depth maps in both cases, but the deblurred image for bottom scene has several artifacts. This scene was captured at close range, and is affected to a greater degree by lens distortions.

scene depth with greater accuracy than other single-shot coded-aperture techniques, while allowing the same quality of deblurring as an un-coded aperture.

Finally, we evaluate depth estimation and deblurring performance on real images captured with the prototype lens. Figures 1 and 6 show the recovered depth maps and all-focus images for three scenes. We are able to recover accurate depth estimates in all cases, and high-quality sharp images for the first two scenes (even in the presence of severe blur in Fig. 1). However, the bottom image in Fig. 6 was captured at close range with the camera roughly at a distance of 40cm from the foreground object. This leads to a greater degree of lens distortion in the induced defocus blur, causing artifacts in the recovered sharp image.

Figure 7 demonstrates the generation of synthetically refocused images from the estimated depth maps and deblurred images, by reassigning each of the depth layers to manually specified defocus levels. Note that despite the poor deblurring quality for the bottom image from Fig. 6, the estimated depth map for that scene still provides opportunities for creating synthetic views where the out-of-focus regions are blurred even further, thus simulating a shallow depth-of-field.



Background in focus

Shallower depth-of-field

Fig. 7. Synthetic refocusing results. These images were generated by reassigning layers from the depth map to user-specified levels of defocus.

7 Discussion

We introduce a single-shot coded-aperture DFD technique that leverages spatio-spectral image statistics. Unlike color-neutral designs for the same task, this approach avoids making a trade-off between depth discriminability and deconvolution ability. The relationship between gradient profiles across different color channels is strong enough to enable accurate, fully-automatic depth estimation, using a light-efficient aperture pattern that does not exaggerate spatial distortion, and this enables efficient, high-quality deblurring. Associated MATLAB code and data are available at `http://vision.seas.harvard.edu/ccap/`.

Worthwhile directions of future work include investigating coded-aperture designs and inference methods that combine the proposed model for gradient-agreement, together with more common greyscale models for sharp images (analysis along these lines was conducted in [21], but was based on a simplifying assumption that the recorded intensities in different channels were exactly equal). While our approach enables depth estimation while matching the deblurring performance of a regular aperture, it might be possible to exceed this performance, for example, by optimizing the spatial aperture pattern for deconvolution [22] (i.e., to induce blur with fewer *zeroes* than a regular aperture), and combining it with spectral filtering for access to depth information.

Furthermore, the image model proposed here, as well as more general spatio-spectral image models [15], are potentially useful in other computational photography applications. For example, coded-exposure techniques could exploit spatio-spectral statistics by introducing a color filter in the light path for a portion of the exposure time, to estimate and invert subject motion blur.

Acknowledgments. The project was supported by NSF award IIS-0926148, ONR award N000140911022, and the US Army Research Laboratory and the US Army Research Office under contract/grant number 54262-CI.

References

1. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Trans. on Graph. (SIGGRAPH). (2007)

2. Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Trans. on Graph. (SIGGRAPH). (2007)
3. Zhou, C., Lin, S., Nayar, S.K.: Coded Aperture Pairs for Depth from Defocus and Defocus Deblurring. Intl. J. Computer Vision. (2011)
4. Krishnan, D., Fergus, R.: Fast image deconvolution using Hyper-Laplacian priors. In: Advances in Neural Info. Process. Sys. (2009)
5. Chaudhuri, S., Rajagopalan, A.N.: Depth from defocus: A real aperture imaging approach. Springer-Verlag, New York. (1999)
6. Hasinoff, S.W., Kutulakos, K.N.: A layer-based restoration framework for variable-aperture photography. In: Intl. Conf. Computer Vision. (2007)
7. Levin, A.: Analyzing depth from coded aperture sets. In: European Conf. Computer Vision. (2010)
8. Cossairt, O., Nayar, S.K.: Spectral Focal Sweep: Extended Depth of Field from Chromatic Aberrations. In: IEEE Intl. Conf. Computational Photography. (2010)
9. Amari, Y., Adelson, E.: Single-eye range estimation by using displaced apertures with color filters. In: Intl. Conf. on Industrial Electronics, Control, Instrumentation, and Automation. (1992)
10. Chang, I.C., Huang, C.L., Hsueh, W.J., Lin, H.C., Chen, C.C., Yeh, Y.H.: Novel 3D handheld camera based on triaperture lens. In: Proceedings of SPIE. (2002)
11. Bando, Y., Chen, B., Nishita, T.: Extracting depth and matte using a color-filtered aperture. ACM Trans. on Graph. (SIGGRAPH Asia). (2008)
12. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM Trans. on Graph. (SIGGRAPH). (2004)
13. Levin, A., Rav Acha, A., Lischinski, D.: Spectral matting. IEEE Trans. Pattern Anal. & Mach. Intell. (2008)
14. Joshi, N., Zitnick, C.L., Szeliski, R., Kriegman, D.J.: Image deblurring and denoising using color priors. In: IEEE Conf. Computer Vision & Pattern Recognition. (2009)
15. Chakrabarti, A., Zickler, T.: Statistics of real-world hyperspectral images. In: IEEE Conf. Computer Vision & Pattern Recognition. (2011)
16. Rajagopalan, A., Chaudhuri, S.: Optimal selection of camera parameters for recovery of depth from defocused images. In: IEEE Conf. Computer Vision & Pattern Recognition. (1997)
17. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. In: ACM Trans. on Graph. (SIGGRAPH). (2006)
18. Gehler, P.V., Rother, C., Blake, A., Minka, T., Sharp, T.: Bayesian color constancy revisited. In: IEEE Conf. Computer Vision & Pattern Recognition. (2008)
19. Boykov, Y., Veksler, O., Zabih, R.: Efficient approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. & Mach. Intell. (2001)
20. Chakrabarti, A., Zickler, T.: Fast deconvolution with color constraints on gradients. Technical Report TR-06-12, Computer Science Group, Harvard University (2012)
21. Baek, J.: Multi-channel coded-aperture photography. Master's thesis, MIT (2008)
22. Zhou, C., Nayar, S.K.: What are Good Apertures for Defocus Deblurring? In: IEEE Intl. Conf. Computational Photography. (2009)