Image and depth retrieval from single monocular images with various point spread functions

Yan Joe Lee
Stanford University
476 Lomita Mall, Stanford, CA
yanjoe@stanford.edu

Jiho Hong
Stanford University
476 Lomita Mall, Stanford, CA
jihohong@stanford.edu

Nayeun Lee
Stanford University
476 Lomita Mall, Stanford, CA
nayeun@stanford.edu

Abstract

Depth estimation of a scene has important applications in fields such as autonomous systems, computer vision, and biological imaging. At the same time, recovery of the all-in-focus image is desirable to extract full information in the scene. The all-in-focus image together with the depth map allows for further post processing such as object detection and refocusing. Here, we explore passive approaches utilizing three different point spread functions (PSF) to 1) construct a depth map, and 2) recover an all-in-focus image from a single monocular image. Specifically, we compare qualitatively and quantitatively the performance of a conventional PSF, a coded aperture PSF, and a double helix PSF.

1. Introduction

When an image of a scene is captured using a conventional camera, the 3-dimensional information of the scene is lost. Furthermore, conventional optical elements used for imaging have a limited depth of field, resulting in blurring of objects too close or too far away from the camera. This loss of information due to blurring is detrimental for object identification, machine vision, and sometimes simply detracts the aesthetics of the captured image.

The recovery of the 3D information of the scene, in particular constructing a depth map, has wide ranging applications including autonomous vehicles, biological imaging, and industrial manufacturing. Different approaches have been explored for depth estimation, which can be broadly divided into active techniques and passive techniques. Active techniques usually employ specialized light sources and scanning systems, for example in lidar systems and structured light imaging. The addition of extra components increases the cost of active techniques, and also makes it more difficult to integrate in compact forms such as mobile phones and miniature robotics.

On the other hand, passive depth estimation relies instead of the ambient light illumination of the scene. Examples include stereo vision and light field cameras. In this project, we explore three passive, monocular techniques for depth estimation from single images using different point spread functions. This approach allows facile integration to conventional imaging systems because the desired point spread function can be achieved using a phase mask or a coded aperture with minimal modification of the optical path. We simulate the image formation and subsequent deconvolution using a conventional (circular) PSF, a coded aperture PSF [1], and a double helix PSF [2].

One of the earlier approaches to depth estimation utilizes focal gradients due to the limited depth of field of conventional optical elements [3]. The amount of defocusing changes with the distance from the point/depth of exact focus. By deconvolving local sub-windows of the image with blur kernels corresponding to different depths, we can choose the best result and hence estimate the local depth. However, it is not easy to determine which deconvolved image (and the corresponding depth scale) is the "best", which led to the development of coded apertures. By properly designing the coded aperture, it can have a frequency response that allows better depth discrimination.

A different approach to estimate depth is by using spatially varying point spread functions. In contrast to both conventional PSFs and coded aperture PSFs that experience a size rescaling with defocusing, spatially variant PSFs can be more sensitive to defocus due to their axial variation. The challenge here is to design physically realizable PSFs using amplitude or phase masks. The double-helix PSF is

an example of a spatially varying PSF which was first used in microscopy [4], and more recently explored for macroscale depth estimation [5]. In the transverse plane, the double-helix has two lobes that rotate with axial defocus, which is shown in Figure 1.

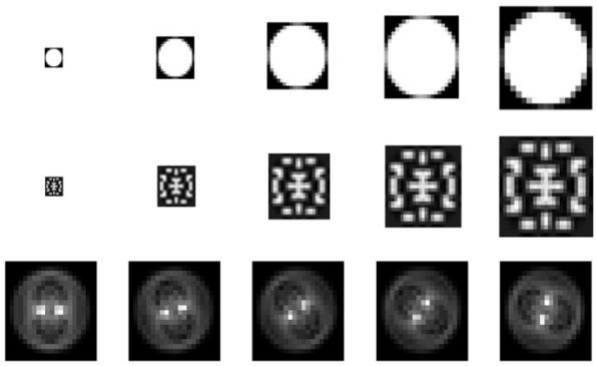

Figure 1. Various kernels at a variety of depths from the focus plane. (Top) PSF for the conventional aperture. (Middle) PSF for the coded aperture[1]. (Bottom) PSF for the double helix.

2. Related Works

In addition to the several approaches mentioned in the introduction, other techniques exist for depth estimation and all-in-focus recovery. Dowski and Cathey [6] have designed an extended-depth-of-field optical system using a cubic phase mask that is insensitive to defocusing. An all-in-focus image can be recovered after postprocessing, but the depth map cannot be reconstructed. A common way to acquire depth information is by using stereo cameras which captures a scene from two different viewpoints [7]. Stereo vision however is limited by the distance between the two cameras used to perform triangulation. This makes it difficult to integrate in compact systems without sacrificing the depth estimation accuracy at large distances. Finally, there are depth estimation methods that rely on taking multiple images of the same scene with multiple focus settings, such as the work of Subbarao and Surya [8].

3. Simulated sensor image formation

We start by simulating a captured sensor image of a 3D scene for various depth-dependent PFSs. For a ground-truth image and a depth map, we use *NYU Depth v2* dataset. We start by discretizing a ground-truth depth map into finite numbers of depth planes. Based on its depth dependence, a PSF at each depth is simulated and precalculated. Then, a blurred sensor image at each depth is also calculated by convolving a ground-truth all-in-focus image with a pre-calculated PSF at a relevant depth. From such a set of blurred sensor images at different depths, we synthesize a depth-dependent sensor image by averaging them with an appropriate weight function. We use each depth plane in a discretized depth map as an occlusion mask. For an all-in-focus image x, a simulated sensor image b is calculated by [9]:

$$b = \sum_{n=1}^{N} (PSF_n * x) \cdot M_n \qquad (1)$$

where $*$ denotes 2D convolution, and $\cdot$ denotes element-wise multiplication. For a set of discrete depth planes (n=1, 2, ⋯, N), PSF_n and M_n represent a PSF and an occlusion mask for the nth depth plane, respectively.

4. Image and depth estimation

Starting from a simulated sensor image, we reconstruct all-in-focus image and depth map. Because both an all-in-focus image and a depth map are unknown, it is challenging to formulate a single optimization problem to estimate them simultaneous. For example, an all-in-focus image cannot be reconstructed directly from simple deconvolution because a blur kernel is not spatially invariant. The spatial variance of a blur kernel is determined by both the depth-dependence of a PSF and the depth map of a 3D scene. To address this, we divide an image and depth estimation problem into three steps: deconvolution, reconstructing an all-in-focus image and a depth map, and smoothing a depth map.

For a set of discrete depths d_n (n=1, 2, ⋯, N) over an estimated depth range, we carry out deblurring of a given sensor image with a PSF at each depth. By selecting each image pixel appropriately among such a set of deblurred images, we reconstruct both an estimated all-in-focus image and an estimated depth map. Each image pixel is selected to minimize a reconstruction error over a local window around its position. Furthermore, we smooth an estimated depth map using a bilateral filter. As we use an estimated all-in-focus image for an intensity kernel, we would obtain the estimated depth map in which depth discontinuities are better aligned with the estimated image.

4.1. Deconvolution with ADMM

To deblur a given sensor image with a PSF at each depth, we carry out regularized deconvolution with the alternating direction method of multipliers (ADMM). Assuming a ground-truth image has sparse gradients like as a natural image, we use a total variation (TV) prior in ADMM. For a captured sensor image b, a deblurred image $\tilde{x}_d$ for a depth of d is expressed as:

$$\tilde{x}_d = \text{minimize}_x \frac{1}{2}\|b - PSF_d * x\|_2^2 + \lambda\|\nabla x\|_1 \qquad (2)$$

where PSF_d is a PSF at a depth of d, and λ represents the strength of the prior. By splitting the objective above into

two independent functions, deconvolution and a TV prior, the TV-regularized deconvolution is formulated as:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|b - PSF_d * x\|_2^2 + \lambda\|z\|_1 \\ &\text{subject to } \nabla x - z = 0 \end{aligned} \quad (3)$$

where z is a slack variable which is only linked to x through the constraint. We solve the augmented Lagrangian of equation (3) through an iterative-update approach.

4.2. Reconstructing image and depth map

From calculated deblurred images with PSFs at different depths, we reconstruct an estimated all-in-focus image and an estimated depth map. For each of discrete depths d_n, we calculate the reconstruction error $e_n = b - PSF_{d_n} * \tilde{x}_{d_n}$ at every image pixel. The average reconstruction error E_n over a local window around the ith image pixel is then defined as [1]:

$$E_n(i) \approx \sum_{j \in W_i} e_n(j)^2 \quad (4)$$

where W_i denotes a local window around the ith image pixel. Based on this, each image pixel of the estimated all-in-focus image is selected to minimize locally the average reconstruction error. At the same time, depth for each image pixel is estimated by the depth of a deblurred image where the all-in-focus image pixel is selected [1]:

$$d(i) = \text{minimize}_{d_n} E_n(i) \quad (5)$$

where $d(i)$ represents the depth for the ith image pixel.

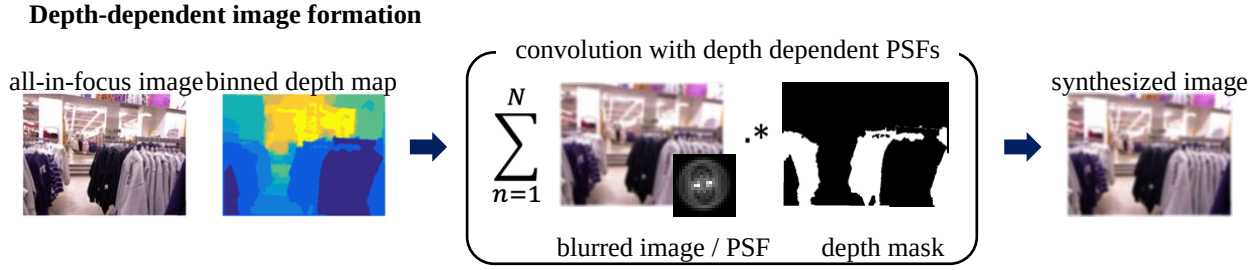

Figure 2. Depth-dependent image formation. (Left) Ground truth image are shown as the all-in-focus image and the binned depth map. (Middle) At each depth steps, convolution of blurred image with PSF at that depth and the depth mask is conducted and summed over to construct the synthesized image (Right).

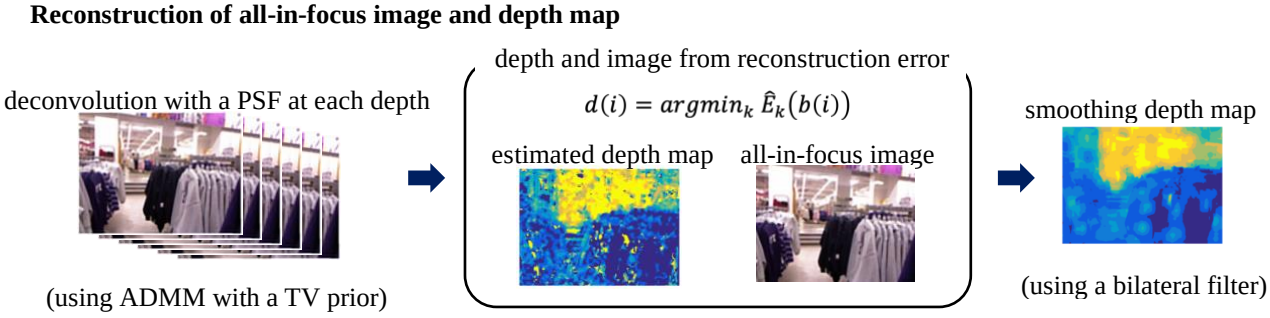

Figure 3. Reconstruction pipeline of all-in-focus image and depth map. (Left) At each depth, images are deconvolved with a PSF for given depth using ADMM with a TV prior. (Middle) Based on the local energy minimum, pixels for the all-in-focus image are selected as well as for the estimated depth map. (Right) Using a bilateral filter, depth map becomes smoother.

4.3. Smoothing depth map

To smooth the estimated depth map, we use a bilateral filter. Note that we use the estimated all-in-focus image for an intensity kernel in our bilateral filter. By applying the bilateral filter with such an intensity kernel, the smoothed depth map would have depth discontinuities aligned to edges of the estimated all-in-focus image. This would allow us to increase the robustness and fidelity of our depth map estimation. For the estimated all-in-focus image $\tilde{x}$, the intensity kernel w_{int} in the bilateral filter is formulated as:

$$w_{\text{int}}(i,j) = \exp\left[-\frac{\|\tilde{x}(j) - \tilde{x}(i)\|_2^2}{2\sigma_{\text{int}}^2}\right] \quad (6)$$

where σ_{int} denotes the standard deviation of the intensity kernel.

5. Results

We show a variety of scenes, recovering both the depth map and fully sharp images. We also compare the performances of conventional, coded aperture and double helix PSF case, using the same deblurring and depth estimation algorithms. In addition, we show some applications made possible by the additional depth information for each image, such as refocusing and scene re-rendering.

5.1. Simulation setup

The original depth map was discretized with 10 steps. To make comparison with the coded aperture [1], we artificially made 13x13 pattern that exactly matches their coded aperture. For the conventional aperture and aperture with double helix PSF, we matched the size of the PSF to that of coded aperture.

Optical model	NYU Depth v2 Image No. 95		NYU Depth v2 Image No. 1	
	PSNR of Focus [dB]	PSNR of Depth [dB]	PSNR of Focus [dB]	PSNR of Depth [dB]
Conventional aperture	21.85	7.91	27.16	6.16
Coded aperture	22.77	14.03	29.05	14.71
Double helix	**23.82**	**18.39**	**29.88**	**17.99**

Table 1. Focus and depth peak signal to noise ratio with different optical models for various datasets. Lowest errors are bolded.

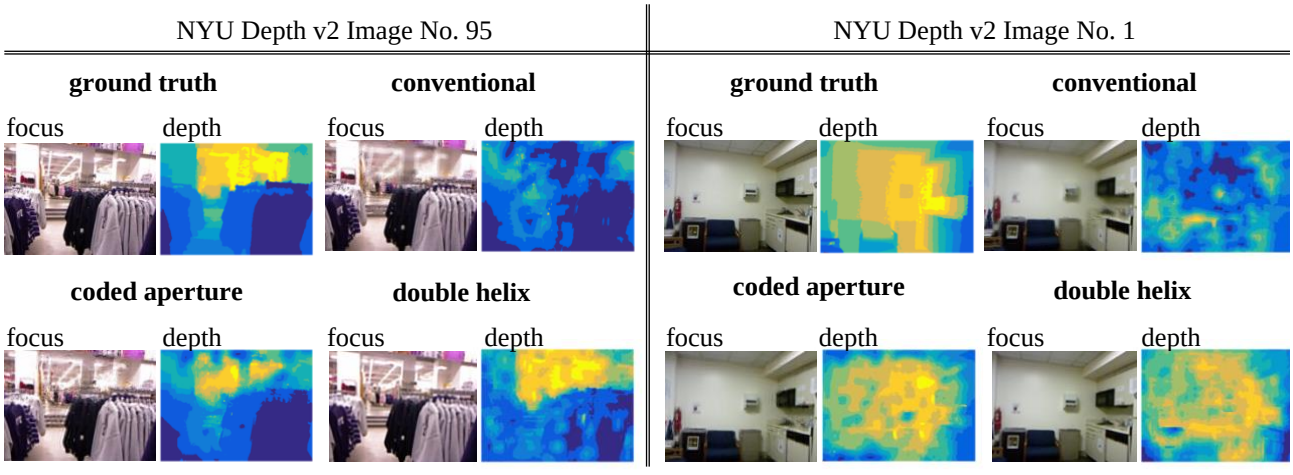

Figure 4. Focus and Depth estimation. (Left) Examples with NYU Depth v2 dataset of Image No. 95 with conventional aperture, coded aperture and double helix point spread function. (Right) Examples with NYU Depth v2 dataset of Image No. 1 with conventional aperture, and coded aperture and double helix point spread function.

For the deconvolution with ADMM with a TV prior, we used our max iteration as 100, λ as 5×10^{-4}, and ρ as 1. When varying these parameters, different results can be produced, and these user-defined parameters are set to be constant over the whole simulation. There are other parameters that can also be changed based on the user's experience such as the window size of reconstruction of local depth map and filter radius and sigma intensity of bilateral filtering for the final depth estimation. However, when the window size is too small, it produces lot of noise, and the window size is large, we lose the details of the depth information.

Due to the ringing pattern of the image edges, depth map and the reconstructed image cannot be trusted in those specific areas. For this reason, we cropped the edge of the final depth map and reconstructed image by the size of given PSF. PSNRs are also calculated after removal of the edges.

5.2. Comparison with existing models

We performed quantitative and qualitative comparison between 3 aperture types on real-world scenes of known depth. Ground truth images were randomly selected in the *NYU Depth v2* datasets. Table 1 and Figure 4 show a summary of simulated results with different aperture types. For each depth map in a single example scene, colors indicate the same depth from the virtual camera. However, there is no relation between colors when we compare both examples.

We observe common trends across all datasets. When using conventional aperture, peak signal to noise ratio (PSNR) values are lowest in the refocused image and recovered depth map. This can be easily understood since our depth estimation algorithm is based on the deblurred image. Using conventional aperture, we can observe the deblurred image shows drastically blurred for the object is far away.

For the depth dependent PSFs, simulated results with aperture with double helix show the best performance both in the focused image and the retrieved depth map. The PSNR is highly dependent on the algorithm parameter we use. We are here showing the results for the given weight of the same priors we used for all image and depth retrieval.

Since we used bilateral filter for the reconstruction final depth map and using artifacts to retrieve the initial depth map, we can observe the depth map is highly dependent on the texture or edges in the scene. For example, in the *NYU Depth v2* Image No. 95, the pillar in the far distance can be retrieved well both in coded aperture and aperture with double helix PSF. This also holds true for the *NYU Depth v2* Image No. 1, where we can easily detect the corners of the rightmost ceiling. Retrieval of the depth information especially of the wall which has the uniform background, it is hard to get useful information out of this.

5.3. Applications

Given the retrieved all-in-focus image and depth map, there is a potential to refocus to certain depth as we have learned from the class, talking about light field camera [10]. This can be useful in the image post-processing, where users can refocus after taking the single image.

6. Discussion and limitations

In this work, we have shown how depth dependent PSFs can be used to retrieve image and depth from single monocular image and compared the performance with conventional aperture. Both coded aperture and even double helix PSF can be utilized for natural scenes, permitting the recovery of both an all-in-focus image and depth from a single image. With the depth dependent PSFs, ringing pattern is intentionally produced after defocus blur which we use to retrieve depth information and all-in-focus image from a single shot. While it is easy to think that conventional aperture with no chromatic aberration will have limited results in doing these tasks, there is clear performance difference among two depth dependent PSFs. This is owing to the fact that simulated double helix PSFs have better distinction along the depth from the focus plane in the given simulation tasks.

Our conclusions are mainly drawn from the relative performance between simulated results. We are not claiming that certain PSF can outperform compared to existing methods. It is clear that there are some limitations. The results are based on our specific deblurring and depth estimation algorithms. The performance can be changed using other algorithms and also using different PSF parameters such as aperture size and the actual depth of field of the scene. As with most classical stereo vision algorithms, this approach is based on the sufficient amount of textures of the scene. Also, the double helix PSFs are challenging to be realized in the real world. They should have long enough focal distance while maintaining spatially varying PSFs. With these reasons, double helix PSFs are currently limited in the super-resolution fluorescent molecule microscope imaging. Nonetheless, if there is proper way to fabricate an aperture with double helix PSFs meeting these criteria, it would be valuable to implement for monocular depth estimation or other visual tasks.

References

[1] Levin, Fergus, Durand, and Freeman. Image and depth from a conventional camera with a coded aperture. ACM transactions on graphics (TOG) 26, no. 3: 70-es, 2007.

[2] Greengard, Schechner, and Piestun. "Depth from diffracted rotation." Optics letters 31, no. 2: 181-183, 2006.

[3] Pentland. "A new sense for depth of field." IEEE transactions on pattern analysis and machine intelligence 4: 523-531, 1987.

[4] Pavani, Sri Rama Prasanna, Michael A. Thompson, Julie S. Biteen, Samuel J. Lord, Na Liu, Robert J. Twieg, Rafael Piestun, and W. E. Moerner. "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function." *Proceedings of the National Academy of Sciences* 106, no. 9 (2009): 2995-2999.

[5] Quirin, Sean, and Rafael Piestun. "Depth estimation and image recovery using broadband, incoherent illumination with engineered point spread functions." *Applied optics* 52, no. 1 (2013): A367-A376.

[6] Dowski, Edward R., and W. Thomas Cathey. "Extended depth of field through wave-front coding." *Applied optics* 34, no. 11 (1995): 1859-1866.

[7] Scharstein, Daniel, and Richard Szeliski. "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms." *International journal of computer vision* 47, no. 1-3 (2002): 7-42.

[8] Subbarao, Murali, and Gopal Surya. "Depth from defocus: A spatial domain approach." International Journal of Computer Vision 13, no. 3 (1994): 271-294.

[9] Chang, Julie, and Wetzstein, Gordon. “Deep optics for monocular depth estimation and 3D object detection.” arXiv:1904.08601v1, (2019)

[10] Ng, Ren, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. "Light field photography with a hand-held plenoptic camera." *Computer Science Technical Report CSTR* 2, no. 11 (2005): 1-11.