Coded aperture to obtain depth and focus image

Moneish Kumar
CMU
moneishk@andrew.cmu.edu

Abstract

Conventional cameras typically produce blurry images for objects that are not in focus. Some camera systems have been developed to either capture all-focus images or extract depth information, but these systems often require specialized hardware and result in reduced spatial resolution. Coded aperture involves modifying a conventional camera to simultaneously capture high-resolution image information and depth information.

1. Introduction

In this study, we present a method for capturing both high resolution RGB images and coarse depth information using a single image capture and minimal modification to a traditional camera system. Our approach involves attaching a simple piece of cardboard to the lens, which may require occasional user assistance. Our method does not utilize any prior information during image generation, but instead employs local deconvolution on regions of interest. This allows for fast image generation with impressive results. Image analysis often involves extracting information from specific regions of interest within the image, rather than the entire image. This is because images often contain few objects of interest, and it is more useful to obtain information about these regions. We focus on retrieving high resolution and depth information for these regions of interest within the image.[need referance].Our approach, which falls under the category of computational photography, differs from other methods in that it allows for the recovery of both image and depth information from a single image. Our method is inspired by techniques in coded aperture imaging and wavefront coding, and involves controlling the defocus produced by the lens to enable the estimation of distance information and the compensation for at least part of the defocus to produce artifact-free images. In contrast to traditional photography, which captures only a 2-dimensional projection of the 3-dimensional world, modifications to recover depth typically require multiple images or active methods with additional equipment such as light emitters. Our system, on the other hand, allows photographers to continue capturing images in the same way they always have while also providing the added benefit of coarse depth information, which can be used for refocusing (extending the depth of field) and depth-based image editing.

Concept To understand the concept of defocus and how we can control and exploit it, 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. However, if the object is placed at a distance Dk away from the focus distance, the rays from the object will land on multiple sensor points, resulting in a blurred image. The pattern of this blur is determined by the aperture cross-section of the lens and is often referred to as a circle of confusion. The amount of defocus, characterized by the blur radius, depends on the distance of the object from the focus plane.

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

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

The goal is to determine a method for recovering both a depth map and a sharp image from a single blurry image. The blurry image (y) is related to the true sharp image (x) through a blur filter (f_k), which is a scaled version of the aperture shape and may also be convolved with the diffraction pattern. The pattern of blur from a conventional lens with the pentagonal disk shape being is formed by the intersecting diaphragm blades. While this type of defocus provides depth cues, they are difficult to exploit because it is challenging to accurately estimate the amount of blur and multiple images are often required.

To address this issue, the researchers explore the possibility of deliberately introducing patterns into the aperture. The captured image will still be blurred as a function of depth, with the blur being a scaled version of the aperture shape. However, the aperture filter can be designed to discriminate between different depths. If the aperture shape is known and fixed, there is only a single unknown parameter

Figure 1. (Left) Image shows the input image, (Right) Image shows the depth map of the scene only for regions of interests. The different color indicate the depths of each of the regions.

Figure 2. (Left) Image shows a cropped version of the input image, (Right) Image shows the focused output notice the plus size and 70 which is sharper in the right image as compared to the image on the left.

(the scale of the blur filter) that relates the blurred image (y) to its sharp version (x).

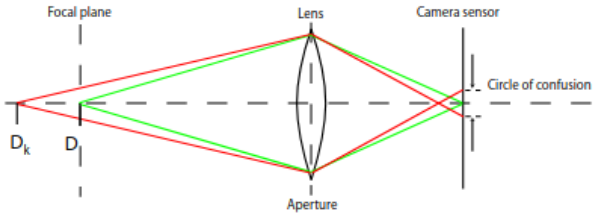

Figure 3. A 2D thin lens model. A point on the focal plane which is at a distance D from the lens is focused at a single point on the sensor whereas a point at a distance D_k from the lens maps to region rather than a point. Figure credits [9]

In real-world scenes, the depth is rarely constant throughout the image. Instead, the scale of the blur in the image (y) varies over its extent, while remaining locally constant. This means that the challenge is to recover not just a single blur scale, but a map of it over the entire image. If this can be reliably achieved, it would have practical utility, as the depth of the scene could be directly computed and the captured image (y) could be decoded to recover the sharp image (x). .

The above discussion only takes into account geometric optics, but a more comprehensive treatment must also consider wave effects, particularly 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. In the case of objects in focus, diffraction dominates, but for defocused areas, the shape of the aperture is most important. As a result, the analysis of defocus usually relies on geometric optics. While the theoretical derivation is based on geometric optics, diffraction in practice by calibrating the blur kernel from real data.

2. Related Work

There are various techniques that can be used to extract depth information from images. Active methods, such as laser scanning [2] and structured light [12][15], involve the use of additional illumination sources to capture 3D information. Passive methods, on the other hand, do not require additional intervention and rely on changes in viewpoint or focus to recover the depth information. Examples of passive methods include stereo imaging, which captures multiple images from different viewpoints [14], and plenoptic cameras [1] [13] [5] [10], which capture multiple viewpoints in a single image at the cost of reduced spatial resolution. Depth from focus and depth from defocus techniques [7] [4] also use multiple images taken from a single viewpoint with different focus settings to estimate the depth. While these methods can produce high-quality depth estimates, they often require multiple images and may not be practical for personal photography.

There have also been attempts to use optical masks [8] [6] to estimate depth from a single image, but these approaches have either not produced high quality images or have only been tested on synthetic images. The goal of the method described in this work is to infer both depth and an image from a single shot without the need for additional user input or loss of image quality.

Another approach to depth estimation is the creation of an all-focus image, which is independent of depth. Wavefront coding [3] is one technique that achieves this by using phase plates to deliberately defocus light rays so that the defocus is the same at all depths, resulting in an image with a large depth of focus but no simultaneous depth estimates. Coded aperture methods, which have been used in astronomy and medical imaging, also allow the collection of more light but do not require the estimation of blur scale as the blur is uniform across the image.

There are several methods for retrieving regions of interest (ROIs) in an image. The advent of machine learning has fast-tracked this task. There are numerous methods using machine learning models to identify and classify specific objects in an image. There are also methods that involve partitioning an image into different segments, with each segment representing a distinct region. Deep neural networks have been widely used for these image segmentation tasks, as they can learn complex patterns in the data and make accurate predictions. One popular approach to image segmentation using deep neural networks is the use of fully convolutional networks (FCN) [11], which are designed to take an input image of any size and output a corresponding segmentation map. FCNs are trained to classify each pixel in the image into one of a set of predefined classes, such as foreground or background. They can also be trained to predict continuous values, such as the probability of a pixel belonging to a certain class. Other popular approaches to image segmentation using deep neural networks include the use of encoder-decoder architectures and attention mechanisms. These methods have achieved state-of-the-art results on a variety of image segmentation tasks, including medical image analysis and autonomous driving.

Overview The structure of this report is as follows: Section 2 describes in detail the process of image capture and post-processing used to obtain the depth map and obtain the high-resolution image. In Section 4 we show the results obtain using this method. We discuss the drawbacks and caveats of this procedure in Section 5. Finally, we look at some of the future directions for this line of work.

3. Method

3.1. Debluring

Given an image and kernel to deconvolve the image with there are numerous methods for deconvolution such as inverse filtering, Weiner deconvolution, constrained least squares, and iterative methods. Levin et. al. [9] suggest a constrained least squares method for deconvolution that relies on heavy priors to obtain the deconvolved image. We however use wiener deconvolution. The main reason for us to use Weiner deconvolution is that it does not have high priors and is more efficient in terms of computing time.

Weiner deconvolution is a method for removing blur from an image that has been distorted by a known blur kernel. It is a type of inverse filtering technique that can be used to restore a clear image from a blurry one. The method involves mathematically reversing the effects of the blur kernel by dividing the blurry image by the kernel, with the goal of reconstructing the original, clear image.

One of the advantages of Weiner deconvolution is that it takes into account the noise present in the image, which can be a major problem with other deconvolution methods. It does this by using a statistical measure called the mean square error (MSE) to determine the optimal balance between removing blur and preserving image detail. The method also allows for the inclusion of a regularization term, which can help to smooth out noise and prevent overfitting.

$$G(f) = \frac{1}{H(f)} \left(\frac{1}{1 + \frac{1}{|H(f)|^2 SNR(f)}} \right)$$

Here, $1/H(f)$ is the inverse of the original system, $SNR(f)$ is the signal-to-noise ratio, and $|H(f)|^2 SNR(f)$ is the ratio of the pure filtered signal to noise spectral density

While Weiner deconvolution can be effective at removing blur from images, it does have some limitations. It is most effective when the blur kernel is known and can be accurately modeled, and it may not work as well when the blur kernel is complex or varies across the image.

3.2. Filter search

After obtaining an image captured using the coded aperture and the corresponding kernels at various scales, the next step is to determine the regions within the image for which the appropriate kernels should be applied in order to deconvolve the image and obtain a high resolution version. Identifying the correct kernel for a particular region also provides the information for the depth of the object in the image. This is the a direct consequence of the concept described in section 1.

Levin et. al. [9] pose their model to give the likelihood of a blurry input image y for a filter f at the scale k for the equation 1. They use an energy estimate to decide the corect filter scale.

Instead of using the criterion previously described, we utilize a simple measure of sharpness to determine the appropriate filter for different regions in the image. According to research by Levin et al. [9], a well-designed coded filter is capable of accurately identifying the scale across multiple scales. Therefore, we use the L1-norm of the gradients of the deblurred images as the criterion for assigning the correct scale.

When an image is captured, if the object in the image is a flat plane at a constant distance from the camera, the blur in the image will be uniform. In this case, restoring the image to its sharp state would involve estimating a single blur scale for the entire image. However, in real-world scenes, there are often variations in depth, which means that a separate blur scale needs to be inferred for each pixel in the image. To address this issue, a practical solution is to use small local windows within which the depth is assumed to be constant. However, if the windows are too small, the depth classification may be unreliable, particularly when the window contains little texture.

To deblur the entire image, we use a set of scaled kernels to generate K possible decoded images. For each scale, we compute the L1-norm of the gradients. A decoded image using a particular scale will usually provide a smooth and plausible reconstruction for parts of the image where that scale is the true scale. However, for areas of the image where the depth differs from the scale being used, the reconstruction will contain serious ringing artifacts and will not be plausible. These artifacts will result in a high L1-norm of the gradients for these areas.

One of the primary benefits of utilizing this criterion is that it allows for the simultaneous deblurring of the input image using all of the available kernels. By applying this method, we can identify those regions of the image that demonstrate the highest level of sharpness across different kernels, which can then be used to determine the appropriate scale for deblurring the entire image. This approach has the advantage of being able to process the deblurring of the input image in parallel, rather than sequentially, potentially leading to a more efficient and effective deblurring process. Figure 4 shows a crop of the input image deblurred using kernels at different scales.

3.3. RoI selection

Using the method described above, we can determine the scale needed to deblur each region in the image. However, this local approach, while effective at capturing a significant amount of information, can also be quite noisy, particularly for regions with uniform, textureless surfaces. To improve

Figure 4. Figure shows the deconvolved image using kernels at different scales. When an incorrect kernel is used to deconvolve it produces artifacts.

the quality of the reconstruction and the estimated depths, we perform the reconstruction process on segments of the image that are identified as regions of interest. We segment different regions of the image using a pre-trained image segmentation model that has been trained on the COCO dataset. The figures 5 and 6 below illustrate the bounding boxes and segmented areas for a given image.

4. Results

The table scene shown in Figure 7 contains the objects at different depths. The depth map, which was obtained provides a fairly accurate reconstruction of the distances from the camera for each of the 3 objects. We see that different scales were selected for each object, the color of the RoI represents the index of the filter chosen for that particular region.

In addition to using segments of the image as regions of interest, we also tried generating images without using any regions of interest, as explained in Section 3.3. The figure in 8 shows the depth estimated using this approach. To deconvolve the input image, we used a 20x20 patch locally with all of the kernels and used a sharpness measure (described in Section 3.2) to assign the scale of the filter. It is important to note that this approach results in noisy depth estimates and also produces incorrect depth estimates for objects that are the farthest away.

5. Failure cases

This method performs resonably well on images with few objects but there are few drawbacks:

- Since the deconvolution doesn't have much prior as such, it results in artifacts as shown in Figure 9
- The method heavily relies on the outputs of the segmentation model. If the regions are incorrectly detected or if some objects are missed then the final output also be incorrect.

References

[1] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence*, 14(2):99–106, 1992. 3

[2] Peter Axelsson. Processing of laser scanner data—algorithms and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54(2-3):138–147, 1999. 3

[3] ER Dowski and W Thomas Cathey. Wavefront coding for detection and estimation with a single-lens incoherent optical system. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2451–2454. IEEE, 1995. 3

Figure 5. Figure shows the bounding boxes for the detection of the region of interest.

Figure 6. Figure shows the segmented regions for the detection of the region of interest.

Figure 7. The figure shows the (left) input and (right) all in focus image of the method.

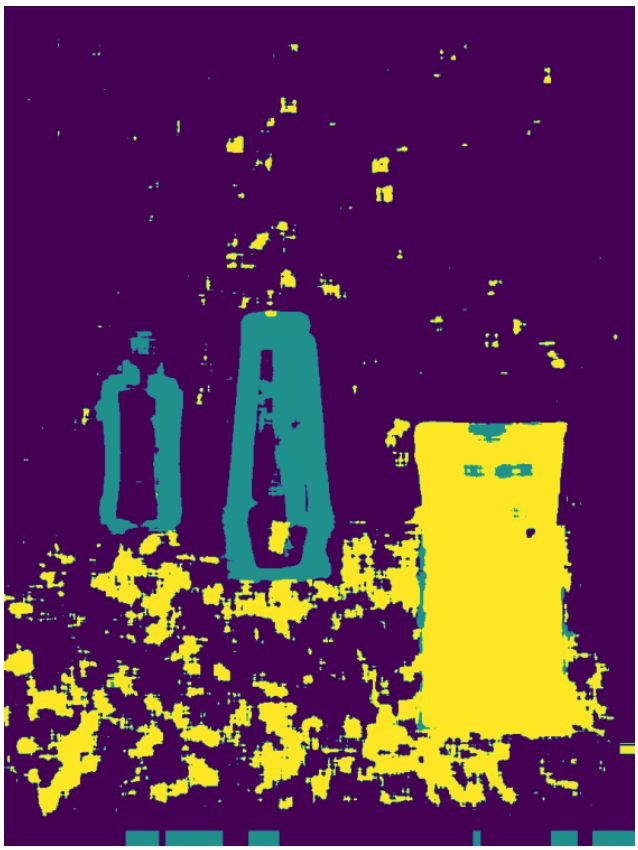

Figure 8. The figure shows the depth estimation without using regions of interest.

[4] Paolo Favaro, Andrea Mennucci, and Stefano Soatto. Observing shape from defocused images. *International Journal of Computer Vision*, 52(1):25–43, 2003. 3

[5] Todor G Georgiev, Ke Colin Zheng, Brian Curless, David Salesin, Shree K Nayar, and Chintan Intwala. Spatio-angular resolution tradeoffs in integral photography. *Rendering Techniques*, 2006(263-272):21, 2006. 3

[6] Adam Greengard, Yoav Y Schechner, and Rafael Piestun. Depth from diffracted rotation. *Optics letters*, 31(2):181–183, 2006. 3

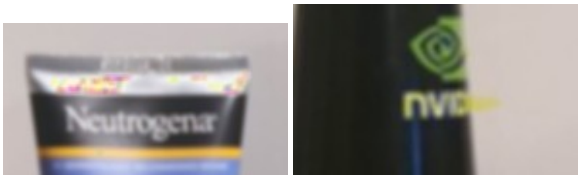

Figure 9. The figure shows the artifacts that are present while doing Weiner deconvolution.

[7] Samuel W Hasinoff and Kiriakos N Kutulakos. Confocal stereo. *International journal of computer vision*, 81(1):82–104, 2009. 3

[8] Shinsaku Hiura and Takashi Matsuyama. Depth measurement by the multi-focus camera. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*, pages 953–959. IEEE, 1998. 3

[9] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70–es, 2007. 3, 4

[10] Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. Light field microscopy. In *ACM SIGGRAPH 2006 Papers*, pages 924–934. 2006. 3

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[12] Shree K Nayar, Masahiro Watanabe, and Minori Noguchi. Real-time focus range sensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(12):1186–1198, 1996. 3

[13] Ren Ng, Marc Levoy, Mathieu Bredif, Gene Duval, Mark Horowitz, and Pat Hanrahan. P.: Stanford tech report ctsr 2005-02 light field photography with a hand-held plenoptic camera. 2005. 3

[14] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002. 3

[15] Li Zhang and Shree Nayar. Projection defocus analysis for scene capture and image display. In *ACM SIGGRAPH 2006 Papers*, pages 907–915. 2006. 3