Depth Estimation Using Segmentation in Natural Images

Shrayan Roy

Supervisor: Dr. Deepayan Sarkar*

Abstract

Estimating depth map from a single photograph is an intriguing problem with many applications. Levin et al.'s proposed a method which requires a modified camera with special coded aperture. We aim to explore whether similar tasks can be done using standard cameras. We have formulated simple parametric models for depth blurring, which we hope will lead to more efficient estimates.

Keywords: AR Prior, Point Spread Function, Depth Map, Image Segmentation

1 Introduction

Traditional photographs are two dimensional projections of a three dimensional scene. The third dimension is Depth, which represents the distance between camera lens and objects in the image. Depth perception is crucial for understanding spatial relationships, with applications in computer vision tasks such as object detection. Additionally, photography and cinematography benefit from depth perception, aiding in the creation of visually compelling compositions. Most modifications to recover depth require multiple images of the same scene [**1**] or active methods with extra apparatus such as light emitters.

Depth estimation based on single image is more challenging problem. Because, we have single observation for each pixel of the image. Depth estimation based on defocus blur exploits the phenomenon where objects appear more blurred depending on their distance from the camera lens, serving as a depth surrogate. Levin et al.'s [**2**] utilized this idea with a sparse gradient prior on natural images to estimate the amount of blur per pixel, albeit necessitating a modified camera with a special coded aperture. Zhu et al.'s [**3**] employed Gabor filters for local frequency component analysis and utilized a simple gradient prior, without the need for a special coded aperture. Recent developments in this field ([**4,5**]) employ deep learning for depth estimation without defocus blur, requiring large image datasets for training. **add blurred images**

*Theoretical Statistics and Mathematics Unit, ISI Delhi < deepayan@isid.ac.in >

In this paper, We will explore whether similar tasks based on depth from defocus can be performed using images from standard cameras, making use of modern high-performance segmentation algorithms such as Segment Anything [**6**]. Our hope is that by starting from a pre-segmented image, it will be easier to estimate the blur kernel without a special coded aperture, using the dependent image gradient prior proposed in [**7**]. In addition to the non-parametric blur kernel estimation method proposed, we have formulated and use simple parametric models for depth blurring, which we hope will lead to more efficient estimates.

2 Mathematical Formulation

When light rays spread from a single point source and hit the camera lens, they either converge to the reference pixel corresponding to the original scene, or they spread out over neighboring pixels. This spreading pattern, determined by the object's distance from the lens or camera movement, is called the *Point Spread Function* (PSF) or *Blur Kernel*. The blurred image is the result of convolving the original sharp image using the PSF. If we assume that the scene remains static for the duration of the photograph and there is no significant camera shake or rotation, then the observed blurred image $\boldsymbol{b}$ of dimension $M \times N$ can be modeled as:

$$\boldsymbol{b} = \boldsymbol{k} \ \otimes \ \boldsymbol{l} \ + \ \boldsymbol{\epsilon}$$

Where $\boldsymbol{k}$ is an $m \times n$ blur kernel or point spread function (PSF), $\boldsymbol{l}$ is the $(M+m) \times (N+n)$ *true latent image* we want to estimate, $\boldsymbol{\epsilon}$ is an $M \times N$ matrix of noise, and $\otimes$ denotes the *convolution* operator.

If we want to reconstruct the original image, we then consider it as an **image deconvolution problem**. Depending on whether the blur kernel is known or unknown, we call it *non-blind deconvolution* or *blind deconvolution* problem. In either case, it is an ill-posed problem because the number of parameters ($(M + m) \times (N + n) + mn$ or $(M + m) \times (N + n)$ depending on *blind* or *non-blind*) is larger than the number of observations MN. One solution to this is to assume some prior for the latent image ($\boldsymbol{l}$). Note that, the model defined above assumes that the associated PSF is *shift invariant*, meaning the same PSF applies to all pixels of the underlying latent image. However, this may not be the case. In the context of depth estimation due to defocus blur, the blur is spatially varying, meaning the PSF function is *not* shift invariant; it is *spatially varying*. Therefore, the above model will not hold in general. We should redefine the above model as:

$$\boldsymbol{b}[\boldsymbol{t}] = (\boldsymbol{k_t} \ \otimes \ \boldsymbol{l})[\boldsymbol{t}] \ + \ \boldsymbol{\epsilon}[\boldsymbol{t}]$$

Where $[t]$ denotes the corresponding elements at pixel location $\boldsymbol{t}$. Note that $t = (t_1, t_2)$ with $(t_1, t_2) \in \{0, 1, ..M-1\} \times \{0, 1, ..N-1\}$. $\boldsymbol{k_t}$ is the spatially varying blur kernel at pixel location

$\boldsymbol{t}$. Now, the problem of estimating blur kernel and latent image becomes more ill-posed because for each pixel, we need to estimate a blur kernel. However, if the blurring is only due to the objects being away from the plane of focus, we can assume special structures of the associated blur kernels. We model the blur kernel $\boldsymbol{k_t}$ as a bivariate probability distribution over a square grid, which aligns with the *Principle of Energy Conservation*. For various pixel locations, we characterize the blur kernel by the parameter $\boldsymbol{\theta_t}$ that determines the scale or spread of the blur kernel. This parameter $\boldsymbol{\theta_t}$ encompasses information regarding the level of blur, hence providing insight into depth. Our objective is to estimate this parameter $\boldsymbol{\theta_t}$ based on the observed blurred image $\boldsymbol{b}$.

3 Priors on Natural Images

We have already noted that the problem of estimating latent image and point spread function (or only latent image) using only the observed blurred image is an ill possed problem and which can be solved if we assume some prior on latent image. But it is not immediately clear what should a proper choice of prior. *Prior elicitation* is an important problem in Bayesian Paradigm and it is important to choose a 'proper prior' which can be used for diverse set of natural images. The prior family used for the latent image is motivated by the observation that when a gradient filter is applied to an image, the distribution of the output has a consistent and distinctive form across a wide range of scene types, with the distribution sharply peaked at zero and relatively heavy-tailed.

This phenomenon is illustrated for **our example images**. We have considered the lag-1 difference along the rows of the image (i.e. *horizontal image gradient*). We can see that distribution of image gradients have a sharper peak at zero and heavier tails than the Gaussian distribution or even Laplace distribution. More frequent values near zero indicates the smooth regions and heavy tails indicates the sharper edges. This observation for natural images is very interesting. It is obvious that, We should choose a prior on image gradients which reflects this observation. Priors with these features are often referred to as *sparse priors* and a useful parametric family to model this is the so called **Hyper-Laplacian Distribution** given by

$$f_\alpha(z) = \frac{\alpha}{2\Gamma(\frac{1}{\alpha})}\exp(-|z|^\alpha), z \in \mathbb{R}$$

With $\alpha > 0$. If $\alpha = 2$ we have Gaussian distribution and for $\alpha = 1$ we have Laplace distribution. Values of $\alpha \in [0.5, 0.8]$ have been found empirically appropriate for natural images. This priors are often computationally difficult to work with. So we use approximate it using scale mixture of zero mean Gaussians (MOG), with

$$\pi(x) = \sum_{j=1}^{J} p_j \frac{1}{\sqrt{2\pi}\sigma_j}\exp(-\frac{x^2}{2\sigma_j^2}), \text{with } \sum_{j=1}^{J} p_j = 1$$

Where p_j's are mixture probabilities and σ_j's are corresponding scale parameters. We refer to this priors as *Sparse MOG Priors*. We incorporate this prior on image gradient using the following modifications to our initial model for blur (**equation number**). If $\delta_h = [-1, 1]$ and $\delta_v = [-1, 1]^T$, then

$$\delta_h \otimes \boldsymbol{b} = \delta_h \otimes (\boldsymbol{k} \otimes \boldsymbol{l}) + (\delta_h \otimes \boldsymbol{\epsilon}) = k \otimes (\delta_h \otimes \boldsymbol{l}) + (\delta_h \otimes \boldsymbol{\epsilon})$$

$$\delta_v \otimes \boldsymbol{b} = \delta_v \otimes (\boldsymbol{k} \otimes \boldsymbol{l}) + (\delta_v \otimes \boldsymbol{\epsilon}) = k \otimes (\delta_v \otimes \boldsymbol{l}) + (\delta_v \otimes \boldsymbol{\epsilon})$$

To keep the notations simple, we will henceforth take the model

$$\boldsymbol{y} = \boldsymbol{k} \otimes \boldsymbol{x} + \boldsymbol{n}$$

Where, $\boldsymbol{y}$ is the horizontal (or, vertical) gradient of observed blurred image. $\boldsymbol{x}$ and $\boldsymbol{n}$ is the same for latent image and noise. For the spatially varying case, it is not immediately clear. We need to assume that the blur kernel is locally constant. i.e. k_t is shift invariant in a neighborhood $\boldsymbol{\eta_t}$ of size $p \times p$ centered at pixel location $\boldsymbol{t}$. This assumption is more or less true, because We expect objects in small local patches to have same depth and hence same level of blur i.e. $\boldsymbol{k_t}$.

$$\boldsymbol{y}[\boldsymbol{t'}] = (\boldsymbol{k_t} \otimes \boldsymbol{x})[\boldsymbol{t'}] + \boldsymbol{n}[\boldsymbol{t'}] \quad \forall \boldsymbol{t'} \in \boldsymbol{\eta_t}$$

So far we have discussed about the marginal distributions of image gradients. But for estimation of latent image or estimation of parameter $\boldsymbol{\theta_t}$ of blur kernel, we need to know the structure of joint distribution of image gradients. Most of the works on Image Processing based Bayesian Image Reconstruction assumes that image gradients are IID with common distribution Sparse MOG Prior. But this assumption is not true as illustrated in the following example. **prior and images**

- Variance calculations

4 Parametric Models for Blur Kernel

To appropriately model the blur kernel, an understanding of certain concepts in *Theoretical Optics* is necessary. Consider the following camera setting diagram: the point F represents the focal point of the lens. Rays traveling parallel to the optical axis of the lens converge to F after diffraction. The distance between the lens center and the focal point is known as the focal length, denoted by *f*. From a point source (pixel), light rays emit in various directions, forming a circular cone-like structure, leading to the concept of the *Blur Circle* or *Circle of Confusion*. Objects at different depths exhibit varying diameters of the blur circle. There exists a well-known relationship between the diameter of the blur circle (c_{diam}) and the depth of objects in a given camera setting.

$$c_{diam} = a_{diam} f \left| \frac{d - d_{focus}}{d(d_{focus} - f)} \right| \approx a_{diam} f \left| \frac{1}{d} - \frac{1}{d_{focus}} \right|$$

Here, d represents the distance of the object from the camera lens (i.e., *Depth*), a_{diam} denotes the aperture diameter, and d_{focus} stands for the distance between the camera lens and the plane of focus. For most cameras, $d_{focus} >> f$, hence the approximation. In a given camera setting (i.e., fixed a_{diam} and f), $c_{diam} \propto \left| \frac{1}{d} - \frac{1}{d_{focus}} \right|$. As we move away from the plane of focus on either side, we encounter a similar type of c_{diam}. Thus, from the diameter of the blur circle, it is challenging to accurately estimate the depth d of an object in an image because for each value of c_{diam}, two possible values of d exist.

From the above discussion, it becomes evident that in the context of blurring caused by depth defocus, the support of the point spread function must be circular rather than square. Additionally, it must exhibit spherical symmetry, indicating that pixels equidistant from the center pixel contribute equally to the convolution. Moreover, the contribution of neighboring pixels in the convolution process should decrease as the distance from the center pixel increases. These phenomena arise due to *diffraction* caused by the camera lens and the boundaries of the circular aperture.