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.

*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 used 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 only converge to the reference pixel corresponding to the original scene, or they spread out over neighboring pixels also. 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:

$$\boldsymbol{b} = \boldsymbol{k} \ \otimes \ \boldsymbol{l} \ + \ \boldsymbol{\epsilon} \tag{1}$$

Where $\boldsymbol{k}$ is an $m \times n$ blur kernel, $\boldsymbol{l}$ is the $(M + m) \times (N + n)$ *true latent image* which we want to estimate, $\boldsymbol{\epsilon}$ is an $M \times N$ matrix of noise, and $\otimes$ denotes the *convolution* operator (By *convolution* we mean *valid* convolution). Constructing the latent image from observed blurred image is called **image deconvolution problem**. Depending on whether the blur kernel is known or unknown, we classify it as *non-blind deconvolution* or *blind deconvolution* problem. In either case, it is an ill-posed problem because the number of parameters is larger than the number of observations MN. One solution to this is to assume a 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 PSF function is *not* shift invariant i.e. it is *spatially varying*. Therefore, (1) will not hold. We redefine it as:

$$\boldsymbol{b}[\boldsymbol{t}] = (\boldsymbol{k_t} \ \otimes \ \boldsymbol{l})[\boldsymbol{t}] \ + \ \boldsymbol{\epsilon}[\boldsymbol{t}] \tag{2}$$

Where $[t]$ indicates 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 some probability density over a square grid. For each pixel location $\boldsymbol{t}$, we characterize the blur kernel by the parameter $\boldsymbol{\theta_t}$ that determines the *scale* or *spread* of the associated probability distribution (e.g. scale parameters in bivariate normal distribution). 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 $\boldsymbol{b}$ for each pixel location $\boldsymbol{t}$.

3 Priors on Natural Images

3.1 Choice of Prior and Basic Ideas

The prior family used for the latent *natural image*[1] is motivated by the observation that when a gradient filter is applied to 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 heavier tails than the Gaussian distribution and Laplace distribution. 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

$$f_\alpha(z) = \frac{\alpha}{2\Gamma(\frac{1}{\alpha})}\exp(-|z|^\alpha), z \in \mathbb{R} \text{ and } \alpha > 0 \tag{3}$$

For $\alpha = 2$ 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. These priors are often computationally difficult to work with. So we approximate it using scale mixture of zero mean Gaussians (MOG), with

$$\pi(x) = \sum_{j=1}^{J} p_j \frac{1}{\sqrt{2\pi}\sigma_j}\exp(-\frac{x^2}{2\sigma_j^2}), \text{with } \sum_{j=1}^{J} p_j = 1 \tag{4}$$

Where p_j's are mixture probabilities and σ_j's are corresponding scale parameters. We refer to this priors as *Sparse MOG Priors*. Levin et al. **[2]** used $\alpha = 0.8$, while Zhu et al. **[3]** utilized $\alpha = 2$, assuming IID gradients. However, Sarkar and Nandy **[7]** demonstrated that the assumption of independent image gradients is incorrect. They have modeled the dependency structure of gradients using a simple first-order AR Model, which we will discuss shortly. Before that, we need to express the image blur model in terms of image gradients. We will do this for (1). If $\delta_h = [-1, 1]$ and $\delta_v = [-1, 1]^T$, then

$$\delta_h \otimes \boldsymbol{b} = \delta_h \otimes (\boldsymbol{k} \otimes \boldsymbol{l}) + (\delta_h \otimes \boldsymbol{\epsilon}) = k \otimes (\delta_h \otimes \boldsymbol{l}) + (\delta_h \otimes \boldsymbol{\epsilon}) \tag{5}$$

[1] By *natural*, we refer to typical scenes captured in amateur digital photography, excluding specialized contexts like astronomy or satellite imaging.

$$\delta_v \otimes \boldsymbol{b} = \delta_v \otimes (\boldsymbol{k} \ \otimes \ \boldsymbol{l}) \ + \ (\delta_v \otimes \boldsymbol{\epsilon}) = k \otimes (\delta_v \ \otimes \ \boldsymbol{l}) \ + \ (\delta_v \otimes \boldsymbol{\epsilon}) \tag{6}$$

To keep the notations simple, we will henceforth take the model by combining (5) and (6)

$$\boldsymbol{y} = \boldsymbol{k} \ \otimes \ \boldsymbol{x} \ + \ \boldsymbol{n} \tag{7}$$

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. By virtue of the *Convolution Theorem*, we rewrite (7) in the frequency domain as

$$\boldsymbol{Y} = \boldsymbol{K} \odot \boldsymbol{X} + \boldsymbol{N} \tag{8}$$

Where, $\boldsymbol{Y}, \boldsymbol{K}, \boldsymbol{X}$ and $\boldsymbol{N}$ are the *Discrete Fourier Transform*'s of $\boldsymbol{y}, \boldsymbol{k}, \boldsymbol{x}$ and $\boldsymbol{n}$ respectively. $\odot$ indicates the *element wise product* operator. $\forall\ \omega = (\omega_1, \omega_2)$ we have

$$\boldsymbol{Y_\omega} = \boldsymbol{K_\omega X_\omega} + \boldsymbol{N_\omega} \tag{9}$$

Remark: In practice the size of the blur kernel $\boldsymbol{k}$ is much smaller compared to the latent image $\boldsymbol{l}$ and $\boldsymbol{x}$. But for (8), the size of $\boldsymbol{K}$ must be the same as the size of $\boldsymbol{X}$. Thus, we pad $\boldsymbol{k}$ symmetrically with zeros to make it the same size as $\boldsymbol{x}$ and then take DFT.

3.2 Motivation and Formulation of Generalized Prior

If the elements of $\boldsymbol{x}$ are zero mean IID random variables with Hyper-Laplacian Distribution. Then by orthogonality of DFT, X_ω's must be IID complex normal when $\alpha = 2$ and uncorrelated and asymptotically complex normal when $\alpha \neq 2$ by CLT of Peligrad and Wu [**8**]. To ascertain whether this is the case, we have plotted the log-transformed modulus of fourier transforms of four sharp images. These can be viewed as realizations from the prior on $\boldsymbol{x}$ or equivalently $\boldsymbol{X}$. There doesn't appear to be any systematic patterns in the plot, suggesting plausibility of the independence assumption. Also, the log-transformed Modulus DFT coefficients are smaller in magnitude at middle frequencies and lighter at large and small frequencies. These findings imply that the variance of DFT coefficients differ across frequencies. To explore further, the **figure** demonstrates smoothing $\log|\boldsymbol{X_\omega}|$ with a Gaussian blur. The resulting pattern is not flat, contrary to what we should expect for IID $|\boldsymbol{X_\omega}|$. It's smaller in the middle and grows towards the corners. This kind of pattern is typically associated with non-IID stationary process.

To check whether the image gradients $\boldsymbol{x}$ are independent, we have plotted the estimated ACF and PACF of the image gradients, both along the direction of the gradient as well as across it. These suggests that image gradients have some local dependence structure, which is stronger across gradient direction. Thus, it is clear that the prior used in most of the works is not entirely

correct. We should model the dependence structure of image gradients in such a way that the above observations with DFT coefficient $\boldsymbol{X_\omega}$'s are consistent. Sarkar & Nandy **[7]** has exactly done this. They have modeled DFT coefficients as non-IID stationary process. In the case of elements of $\boldsymbol{x}$ being IID random variables with variance σ^2, the matrix of DFT coefficients also have constant variance σ^2 (By Orthogonality of DFT). If elements of $\boldsymbol{x}$ are not IID, DFT coefficients have no longer same variance but they are still independent. We introduce the model for *Generalized Prior* as follows

Model: *The process $\boldsymbol{x}$ can be transformed to another process $\boldsymbol{z}$ with IID components with distribution (4) by the following relation: $X_\omega = \sqrt{g_\omega} Z_\omega$, i.e. $Var(\boldsymbol{X_\omega}) = g_\omega \sigma^2$ and $\boldsymbol{Z}$ is the DFT of $\boldsymbol{z}$.*

If we assume the elements of $\boldsymbol{x}$ are marginally distributed as (4) with correlation structure given by (10), then we can easily find an explicit expression for g_ω. For the rest of the paper we will work under this simple generalized lag-1 auto regressive model. ρ_1 and ρ_2 are hyperparameters and values of $\rho_1 = 0.3$ and $\rho_2 = 0.6$ are found to be appropriate for most of the images. Note that in this case, $g_\omega = 1$ we have IID *Sparse MOG Prior*.

$$\text{Corr}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{kl}) = \rho(\boldsymbol{x}_{ij}, \boldsymbol{x}_{kl}) = \rho_1^{|i-k|} \rho_2^{|j-l|} \tag{10}$$

For spatially varying blur, it is not immediately clear how we can express (2) in terms of image gradients. We need to assume that the blur kernel $\boldsymbol{k_t}$ is shift invariant in a neighborhood $\boldsymbol{\eta_t}$ of size $p_1(\boldsymbol{t}) \times p_2(\boldsymbol{t})$ containing $\boldsymbol{t}$ (11). 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. Then, we apply the prior discussed above to that specific patch of the image.

$$\boldsymbol{y}[\boldsymbol{t'}] = (\boldsymbol{k_t} \otimes \boldsymbol{x})[\boldsymbol{t'}] + \boldsymbol{n}[\boldsymbol{t'}] \quad \forall \boldsymbol{t'} \in \boldsymbol{\eta_t} \tag{11}$$

4 Parametric Models for Blur Kernel

From a single point source, light rays emit in different directions and fall on the lens of the camera. The diffracted rays by the lens then form a circular shape on the camera sensor plane, which is called the *Circle of Confusion* or *Blur Circle*. The diameter of the blur circle (c_{diam}) and the depth of an object in a given camera setting are related as:

$$c_{diam} = a_{diam} f \left| \frac{d - d_{focus}}{d(d_{focus} - f)} \right| \approx a_{diam} f \left| \frac{1}{d_{focus}} - \frac{1}{d} \right| \tag{12}$$

In a given camera setting (i.e., fixed a_{diam} and f)[2], $c_{diam} \propto \left| \frac{1}{d_{focus}} - \frac{1}{d} \right|$. As we move away from the plane of focus on either side, we encounter a similar type of c_{diam}. Thus, from the diameter

[2]For most cameras, $d_{focus} >> f$, hence the approximation.

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. It is evident from (12) that, the support of PSF in our case must be circular rather than square. Due to *diffraction* caused by the camera lens and the boundaries of the circular aperture, we expect the intensity distribution of light to be spherical symmetrically distributed over the circular support. Keeping all these in mind we propose the following models for blur kernel.

- **Disc Kernel:** It is the simplest model for blur kernel. The assumption being that uniform spread of light over disc. We characterize the kernel using the parameter r.

$$k(x,y) = \frac{1}{\pi r^2} \times \mathrm{I}_{\{x^2+y^2 \,\leq\, r^2\}}$$

- **Circular Gaussian Kernel:** An obvious choice of kernel in any scientific study is the the widely used Gaussian/Normal distribution. Here, we will consider a truncated version of it over circular region. We characterize the kernel using the radius of circle r and scale parameter h.

$$k(x,y) = \frac{1}{2\pi h^2} e^{-\frac{x^2+y^2}{2h^2}} \times \mathrm{I}_{\{x^2+y^2 \,\leq\, r^2\}}$$

- **Circular Cauchy Kernel:** We know that the intercept on the x-axis of a beam of light coming from the point $(0,h)$ under certain assumption is distributed as a Cauchy$(0,h)$ distribution. Extending this concept to the two-dimensional case, we have a bivariate Cauchy distribution over a circular support. We characterize this using the radius of the circle r and the scale parameter h.

$$k(x,y) = \frac{h}{2\pi} \frac{1}{(x^2+y^2+h^2)^{3/2}} \times \mathrm{I}_{\{x^2+y^2 \,\leq\, r^2\}}$$

- **Rectangular Gaussian Kernel:** We can also consider a truncated Gaussian kernel defined over a finite square grid S_h, characterized by h. In this scenario, we have only one parameter h.

$$k(x,y) = \frac{1}{2\pi h^2} e^{-\frac{x^2+y^2}{2h^2}} \mathrm{I}_{\{(x,y)\in S_h\}}$$

Remark: The parameter r in both the Circular Gaussian and Circular Cauchy kernel characterizes the radius of the blur circle (i.e., $c_{diam}/2$). For a given camera setting, there exists a relation between h and r, namely $h = \kappa \times c_{diam}$ with $\kappa \in (0, \frac{1}{2}]$ and depends upon particular camera. Which means that we cannot change h and r independently.

5 Maximum Likelihood Estimation of Blur Kernel Parameters

Our main focus in this section is to develop *Maximum Likelihood Estimation* procedure for the parameter $\boldsymbol{\theta_t}$ of blur kernel for each pixel location $\boldsymbol{t}$. For the first three kernels $\boldsymbol{\theta_t} = (r_t, h_t)$, and for the Gaussian kernel $\boldsymbol{\theta_t} = h_t$. Our objective is to estimate these parameters based on the observed image $\boldsymbol{b}$ or equivalently $\boldsymbol{y}$. For that we need joint distribution of elements of $\boldsymbol{y}$.

For easier calculations we move to frequency domain. In the case of generalized prior we have $\boldsymbol{X_\omega} \sim \mathcal{CN}(0, \sigma^2 g_\omega) \ \ \forall \omega$ independently. In addition, if the errors in the original image, given by $\boldsymbol{\epsilon}$, are assumed to be IID Gaussian, then its gradient $\boldsymbol{n}$ will have correlated elements and successive elements in the direction of the gradient will have correlation 0.5, while all other pairs will be uncorrelated. This will induce a non-constant variance for $\boldsymbol{N_\omega}$, given by a function h_ω such that $Var(\boldsymbol{N_\omega}) = \eta^2 h_\omega$. Hence, $\boldsymbol{N_\omega} \sim \mathcal{CN}(0, \eta^2 h_\omega)$ independently. An explicit expression for h_ω can be found similarly to that of g_ω.

From (9) we have, $\boldsymbol{Y_\omega} \sim \mathcal{CN}(0, \sigma^2 |K_\omega|^2 g_\omega + \eta^2 h_\omega) \ \ \forall \boldsymbol{\omega}$. For a given choice of model for blur kernel, we estimate parameters $\boldsymbol{\theta}$ based on $\boldsymbol{Y_\omega}$'s. For ease of calculation we find the likelihood based on $|\boldsymbol{Y_\omega}|^2$'s.

Result: *If* $Z \sim \mathcal{CN}(0, \sigma^2)$. *Then,* $Re(Z)$ *and* $Im(Z)$ *follows* $\mathcal{N}(0, \frac{\sigma^2}{2})$ *independently. Hence,* $|Z|^2 = Re^2(Z) + Im^2(Z) \sim \frac{\sigma^2}{2}\chi^2_2 \equiv Exp(\lambda = \frac{1}{\sigma^2})$.

Using the above result, $\boldsymbol{|Y_\omega|^2} \sim \text{Exp}(\lambda_\omega = \frac{1}{\sigma^2 |K_\omega|^2 g_\omega + \eta^2 h_\omega}) \ \forall \boldsymbol{\omega}$ independently. If $f_\theta(|Y_\omega|^2)$ denotes the pdf of $\text{Exp}(\lambda_\omega = \frac{1}{\eta^2 h_\omega + \sigma^2 |K_\omega|^2 g_\omega})$ for given parameters $\boldsymbol{\theta}$ (say,) of blur kernel. Then, likelihood of $\boldsymbol{|Y_\omega|^2}$ is given by -

$$f_\theta(|Y_\omega|^2, \forall \omega) = \prod_\omega f_\theta(|Y_\omega|^2) \tag{13}$$

The above considerations are only for gradients in a particular direction, i.e., *horizontal* or *vertical*. We should incorporate both directions in the estimation procedure. For simplicity we assume that they are independent although clearly they are not as they must integrate to the same latent image $\boldsymbol{l}$. The joint likelihood is given by

$$L(\theta) = L_h(\theta) \times L_v(\theta) = f_\theta(|Y_{h,\omega}|^2, \forall \omega) \times f_\theta(|Y_{v,\omega'}|^2, \forall \omega') \tag{14}$$

Our goal is to find maximizer of $L(\boldsymbol{\theta})$ or equivalently of $\log(L(\boldsymbol{\theta})) = log(L_h(\boldsymbol{\theta})) + log(L_v(\boldsymbol{\theta}))$ as a function of $\boldsymbol{\theta}$. For the spatially varying case, we simply apply this procedure to local patches $\boldsymbol{\eta_t}$ for all pixel locations $\boldsymbol{t}$ in the image domain.

6 References and Bibliography

7 Appendix