Use of Linear Diffusion in Depth Estimation Based on Defocus Cue

Vinay P. Namboodiri and Subhasis Chaudhuri

Department of Electrical Engineering, Indian Institute of Technology, Bombay

Mumbai 400076. India.

{vinaypn, sc} @ ee.iitb.ac.in

Abstract

Diffusion has been used extensively in computer vision. Most common applications of diffusion have been in low level vision problems like segmentation and edge detection. In this paper a novel application of the linear diffusion principle is made for the estimation of depth using the properties of the real aperture imaging system. The method uses two defocused images of a scene and the lens parameter setting as input and estimates the depth in the scene, and also generates the corresponding fully focused equivalent pin-hole image. The algorithm described here also brings out the equivalence of the two modalities, viz. depth from focus and depth from defocus for structure recovery.

1 Introduction

The idea of diffusion has been one of the important methodologies in the field of computer vision. It stems largely from the idea of modeling the image (observation) generation process using the heat equation. The pioneering work was done by Witkin in [14] where he proposed a scale space for images based on smoothing of images using a Gaussian kernel. Koenderink in [2] proved that this was equivalent to solving the heat equation. This approach has subsequently been widely used in low level vision tasks like segmentation and edge detection.

In this paper we discuss how the linear diffusion principle can be used for depth estimation based on defocus as the cue. In depth estimation using defocus as the cue, the basic principle is to use the characteristics of the imaging system. There have been two methodologies followed, one is to obtain depth from focus [3] and the other to obtain depth from defocus [1].

1.1 Depth from Focus

In the procedure for obtaining depth information from focus, a sequence of images of a scene is obtained by continuously varying the distance between the lens and the image detector [12]. Then for each pixel in the image its corresponding fully focused observation is estimated from the sequence of images. From the fully focused image point the distance of the corresponding object point is calculated using the standard lens equation $1/f = 1/u + 1/v$ where f is the focal length, u is the distance of the object from the lens plane and v is the distance of the focused image from the lens plane.

1.2 Depth from Defocus

When a point light source is in focus, all light rays that are radiated by the object point and intercepted by the lens, converge at a point on the image plane. When the point light source is not in focus, its image on the image plane is not a point, but a circular patch resulting in a blur as can be seen from fig (1). In depth from defocus, given two images of a scene recorded with different camera settings, one obtains estimates of the blur at each point [1]. Subsequently, by using an estimate of the blur, one can recover the depth information in the scene with the knowledge of the lens parameters.
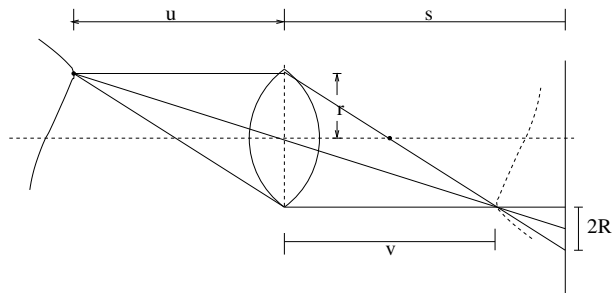

Figure 1: Illustration of image formation in a convex lens.

In this paper we show that, given two observations obtained by two sets of lens parameters, as is commonly employed in depth from defocus technique, using the diffusion equation we can generate the entire set of images in the *defocus space* of the input images. The defocus space of a scene refers to the continuous space of all possible observations obtainable by varying the lens parameters in between those two lens settings. This concept is further elucidated in section 3.3. In this method, the defocus blur is not explicitly calculated as it is done in depth from defocus techniques. Instead, by using diffusion, for each pixel we can obtain the corresponding fully focused observation in the defocus

space. Then using that observation and the corresponding virtual lens parameters we can recover the depth information from the lens equation. As a by-product, we also obtain the fully focused pin-hole image from two defocused observation. The diffusion process simulates the depth from focus technique by generating images in the defocus space of the observation. Many separate observations as required for the depth from focus technique are no longer required. Infact, using the diffusion technique, the two modalities of estimation of depth can be considered to be equivalent. This is discussed further in section 3.4.

In the next section we give a brief overview of the related work done. In section 3 we outline the theoretical basis for the formation of the defocus image space of an observation based on the diffusion process. In section 4 we present the basic algorithm for depth estimation using diffusion. In section 5 we analyse the procedure and consider the practical issues involved in the implementation of this method. In section 6 we present the experimental results obtained. We conclude the paper in section 7.

2 Related Work

2.1 Depth From Focus

There are a number of papers in the literature which address the problem of obtaining depth information from focus. This includes work by Nayar and Nakagawa [7], and work by Subbarao and Choi [12]. The basic method followed has been to obtain different focus levels by adjusting the camera parameters, i.e. either the lens to image plane distance v, the focal length f or the aperture radius r. The methods involve obtaining many observations for the various camera parameters and estimating the focus measure using various criterion functions. Krotkov [3] has experimentally evaluated several such criterion function including the Laplacian and Teningrad operators. The fundamental weakness of this method is, however, the time required for image acquisition. In practice about ten or even more images are required to estimate the depth of a scene for a reasonable level of accuracy.

2.2 Depth From Defocus

The research on depth from defocus (DFD) was introduced by Pentland [9]. He identified the problem of DFD as an estimation of linear space variant blur. The defocus parameter was recovered using the deconvolution in the frequency domain. However, the method depended on the availability of a perfectly focused image of the scene as one of the observations. Subbarao [11] proposed a more general method in which he removed the constraint of one image being formed with the pinhole aperture. In [1], the authors have posed DFD as a problem of regularized space variant blur identification. They provide various methods for solving this problem by modeling the depth/image as an Markov Random Field (MRF). More work in this area has been reported by Favaro *et al.* [8].

2.3 Heat Equation and Diffusion

In this paper we use the technique of diffusion for synthesizing new virtual observations in the defocus space. The idea of diffusion can be traced to that of scale space filtering by Witkin [14]. Koenderink [2] showed that this is equivalent to solving the heat equation. This scale space approach was extended by Perona and Malik in their landmark paper [10] where they proposed a nonlinear scale space model, aimed at preserving important features such as edges. The model changes its behaviour based on the conduction coefficient associated in a region of an image and achieves forward diffusion in the interior region and at the boundaries it acts in the opposite direction. In general the inverse diffusion approach can be thought of as reversing the heat equation. This reverse heat equation is however ill-posed and there has been a substantial amount of work done for stabilizing the reverse heat equation. Rudin, Osher and Fatemi in [4] introduced the "shock filter" where they proposed a pseudoinverse, where the inverse diffusion propagation term is tuned by the sign of the laplacian. There has been a lot of research done along similar lines where various nonlinear inverse diffusion models have been proposed. In linear scale space theory, recently interesting work has been done by Lindeberg [5], where he provides a theoretical analysis of the linear scale space theory and also observes that Gaussian and higher orders of the Gaussian kernel are the only admissible kernels based on the admissibility conditions for linear scale space.

3 Formation of Defocus Space

3.1 Diffusion Process

Consider the classical equation for the isotropic diffusion of heat which is given by the following partial differential equation:

$$\frac{\partial I(x;t)}{\partial t} = a\left(\frac{\partial^2 I(x;t)}{\partial x^2}\right) \qquad (1)$$

Here the constant a is the thermometric conductivity or diffusivity [13]. The equation above describes how heat diffuses over a surface, given an initial temperature distribution with time. It is assumed here that the diffusion of heat is uniform in all directions. Note that the variable x in eqn(1) represents the space and can be considered a 2D function. Consider that $I(x, t = 0)$ is an image. The solution of the heat equation can be obtained in terms of convolution of the image with a temporally evolving Gaussian kernel [6]. This is known as the source solution for the heat equation [13] and is given by $\sigma^2 = 2at$ where σ denotes the spread of the Gaussian kernel used through out in this paper. As the

image is progressively convolved with a Gaussian kernel, it gets increasingly more blurred thereby representing the image information at a different scale. Note that as $t \to \infty$ this corresponds to a fully diffused image. This is the basic idea underlying scale space analysis. Also note that the process is not defined for $t < 0$, a fact that will be utilized later to define the *extended defocus space*.

3.2 Basic Model of Defocus

Consider the image formation process in a real aperture camera employing a thin lens [1]. When a point light source is in focus, all light rays that are radiated from the object point and intercepted by the lens converge at a point on the image plane. When the point is not in focus, its image on the image plane is no longer a point but a circular patch of radius that defines the amount of defocus associated with the depth of the point in the scene. It can be shown that

$$\sigma = \kappa r v (\frac{1}{F} - \frac{1}{v} - \frac{1}{Z})$$ (2)

where r is the radius of the aperture, v is the lens-to-image plane distance, F is the focal length of the lens, Z is the depth at that point and κ is a camera constant that depends on the sampling resolution on the image plane. Let $I(x, y)$ be the pin-hole image of the scene. From the eqn(2) we note that $C = (r, F, v)$ defines the camera parameters each of which may be changed to effect a different amount of defocus blur for a fixed depth.

The depth related defocus process is linear but not space invariant. Assuming a diffraction-limited lens system and a constant depth in the scene (this assumption will be relaxed at a later stage), the point spread function of the camera system at a point (x, y) may be approximately modeled as a circularly symmetric 2-D Gaussian function [1]:

$$h(x, y) = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2 + y^2}{2\sigma^2})$$ (3)

where the blur parameter σ is obtained from eqn(2). Assuming the depth to be constant everywhere, the observed defocussed image $E(x, y)$ is given by

$$E(x, y) = I(x, y) * h(x, y).$$ (4)

This equation can be directly related to the solution of the diffusion equation in terms of the Gaussian as discussed in section 3.1. The real aperture imaging can thus be thought of as providing a real world example of scale space theory. The eqn(4) can be represented by taking its Fourier transform. Denoting the Fourier transform of a function $f(x, y)$ by $\hat{f}(\omega_x, \omega_y)$ we obtain

$$\hat{E}(\omega_x, \omega_y) = \hat{I}(\omega_x, \omega_y)\hat{h}(\omega_x, \omega_y)$$
$$= \hat{I}(\omega_x, \omega_y)exp(-\frac{\sigma^2(\omega_x^2 + \omega_y^2)}{2})$$ (5)

3.3 Defocus Space

For a given scene, one can have two defocused observations E_1 and E_2 corresponding to two different camera parameter settings C_1 and C_2, such that the resulting blur parameters are σ_1 and σ_2, assuming $\sigma_1 > \sigma_2$ without loss of generality. For the two observations E_1 and E_2, a defocus space can be defined.

Definition 1: *Defocus space*

> *The defocus space is defined to be the set of all possible observations E for a given scene generated by varying the blur σ as a combination of the associated blur parameter σ_1 and σ_2 in the two observations E_1 and E_2 respectively, by the following relation*

$$\sigma^2 = \alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2$$ (6)

> *for all values of $0 \le \alpha \le 1$.*

This is equivalent to generating $I(x, y, t)$ for $t_1 \le t \le t_2$ given the states $I(x, y, t_1)$ and $I(x, y, t_2)$ at two specified time instants t_1 and t_2 in the heat diffusion eqn(1). Substituting eqn(6) in eqn(5) we obtain:

$$\hat{E}(\omega_x, \omega_y) = \hat{I}(\omega_x, \omega_y)$$
$$exp[-\frac{1}{2}(\alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2)(\omega_x^2 + \omega_y^2)]$$
$$= \{\hat{I}(\omega_x, \omega_y)exp[-\frac{\sigma_1^2(\omega_x^2 + \omega_y^2)}{2}]\}^\alpha$$
$$\{\hat{I}(\omega_x, \omega_y)exp[-\frac{\sigma_2^2(\omega_x^2 + \omega_y^2)}{2}]\}^{1-\alpha}$$

or $$\hat{E}(\omega_x, \omega_y) = \hat{E}_1^\alpha(\omega_x, \omega_y)\hat{E}_2^{(1-\alpha)}(\omega_x, \omega_y).$$ (7)

The relation given in eqn(7) is equivalent to the notion of scale space as formed by the diffusion equation. This can be noticed as eqn(7) can be thought of as convolving the image $I(x, y)$ with a time varying Gaussian kernel. This is because convolving a Gaussian function with another Gaussian function always results in a Gaussian function. The eqn(7) effectively reduces to convolving the original image $I(x, y)$ with a Gaussian kernel which varies with time (in this case α) according to the relation given in eqn(6).

The defocus blur σ could be present physically due to any of the following camera parameters-aperture, the lens to image plane distance, the focal length or even a combination of these, as shown in eqn(2). A monotonic variation in any of the lens parameters can generally result in a non-monotonic variation in the blur (for instance as v is changed from an initial value, σ reduces, becomes zero and then subsequently increases), signifying both sides of the defocus cone (see Fig.2 for illustration). The diffusion based defocus space generation process however generates the blur in a monotonic manner, i.e we are restricted to one side of the

defocus cone. By continuously varying the parameter α, we can generate any virtual observation for defocus setting lying between the lines AB and CD in Fig.2 using the eqn(7). The defocus space thus consists of all possible observations of the defocus blur $\sigma_1^2 \leq \sigma^2 \leq \sigma_2^2$.

3.4 Equivalence of DFF and DFD

Corresponding to the notion of continuous defocus space as introduced in the previous section, a practical counterpart of this defocus space would be a sampled defocus space. This corresponds to generating the defocus space for discrete values of α between 0 and 1. In Fig.2, the lines corresponding to $A_1B_1, A_2B_2, ...A_nB_n$ may represent one such possible set of sampled defocus space. The sampled defocus space generated for an image is similar to the physically obtained focused image space described in [12].

So far we have considered a restricted range of α between $[0, 1]$. Now we relax this condition and something interesting happens. If the values of α beyond the range $[0, 1]$ is considered then the defocus space generated is the extended defocus space.

Definition 2: *Extended defocus space*

The extended defocus space is defined to be the set of all possible observations E for a given scene generated by varying the blur σ as a combination of the associated blur parameter σ_1 and σ_2 in the two observations E_1 and E_2 respectively, by the following relation

$$\sigma^2 = \alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2 \qquad (8)$$

for all values of $\beta \leq \alpha \leq \infty$.

Here the value of β is the value of α such that $\sigma^2 = 0$ in eqn(8), i.e. resulting in a fully focused observation. This can be obtained from the diffusion equation since corresponding to the image $I(x, y, t)$ with $t \rightarrow \infty$ we can obtain an observation $E(x, y)$ with $\alpha \rightarrow \infty$. This represents the fully diffused image. Similarly for each point there exists a value $\alpha = \beta < 0$ corresponding to $t = 0$. This corresponds to a fully focused observation, i.e. $\sigma^2 = 0$. Thus the extended defocus space is defined for the range $\alpha \in [\beta, \infty)$. In the range $\alpha = [\beta, 0]$ the process, instead of being diffusion becomes an inverse diffusion. Beyond this range, the defocus space is undefined since one cannot have the blur $\sigma^2 < 0$. This is illustrated in the Fig.2. Depth from defocus(DFD) methodology estimates the space variant blur whereas depth from focus (DFF) methodology estimates the focused image point. It is possible to use the techniques in DFD methodology to estimate the space variant blur using just two observations, whereas DFF requires many samples to estimate the fully focused point. Here as we have shown, it is possible to generate the extended defocus space for the image using just two observations. Thus

both the techniques can be considered fundamentally equivalent, rendering the need for multiple samples to be redundant. This diffusion based process thus provides an equivalent means for estimating the depth from the known lens parameters using either depth from defocus or depth from focus.
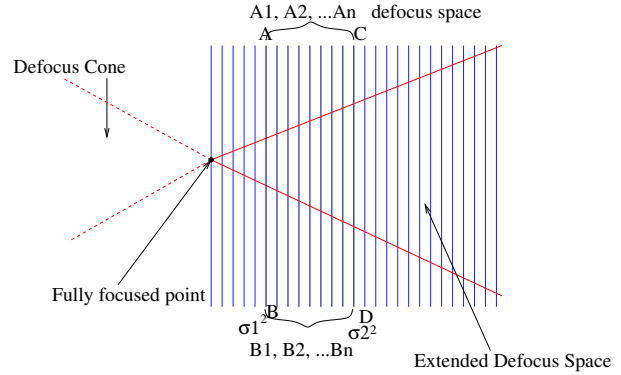

Figure 2: Illustration of the concept of defocus space for a particular scene

4 Algorithm for Depth Estimation

The derivation of eqn(7) is based on the assumption of constant depth. When there is depth variation in the scene, eqn(7) is no longer valid as the blurring process becomes shift variant, implying a non-homogeneous diffusion process. This corresponds to the following diffusion equation

$$\frac{\partial I(x, y; t)}{\partial t} = a(x, y)\left(\frac{\partial^2 I(x, y; t)}{\partial x^2} + \frac{\partial^2 I(x, y; t)}{\partial y^2}\right) \quad (9)$$

Here a is no longer a constant but is now a function $a(x, y)$ and this is handled by forming a small MxM window about a point over which the depth can be assumed to be constant as is done commonly in all literaure. Using this modification the defocus space for a scene can be created locally even in the depth varying case. The depth estimation is done by obtaining the fully focused point for each image. The process of creating the defocus space is a monotonic process. As α varies, the characteristics of the process changes from diffusion to inverse diffusion and the deblurring of the defocused observations happen. In obtaining the fully focused image the value of α is not restricted to lie between 0 and 1, rather we go for values of $\alpha < 0$. The characteristic of the convolution changes from a low pass filter to a high pass filter for $\alpha < \beta$. The defocus process has to be stopped when the fully focused point is reached. This stopping point is estimated empirically from the virtually created observations using a band pass filter, similar to the way it is done in DFF methods, albeit with a difference as in DFF various kinds of high pass filters are used

since the observations in DFF after the fully focused observation are again blurred observations. The various steps of the algorithm are as follows:

1. Divide the observed images E_1 and E_2 into overlapping MxM windowed representations.

2. Obtain the FFT of the corresponding windows in E_1 and E_2.

3. Synthesize a sample of the defocus space corresponding to a particular value of $\alpha \in [\beta, 0)$ for each window using eqn(7). Note that β is unknown as the value of β would give us the depth.

4. Estimate the amount of defocus using a defocus criterion function which is essentially a band pass filter and detect whether a fully focused point is reached.

5. Using the corresponding virtual lens parameters, calculate the value of depth at the point.

This algorithm is sequentially executed for all pixels in the image till the corresponding pin-hole observation of the scene is obtained and a dense depth map is generated.

5 Computational Difficulties

The algorithm uses the windowed Fourier transform. In some cases, especially, where the grey level variance in the window values is very low, signifying a textureless scene, there might be a problem as the spectral components may be nearly 0. When the value of α goes beyond 0 and 1 range, then potentially a division by zero error can occur in eqn(7). This can be avoided by marking such windows as being out of computation. Mathematically it signifies that the depth cannot be estimated for homogeneous regions.

Another factor which adversely affects the method is its sensitivity to quantization error. Generally, 8 bit quantization of the scene results in very noisy virtual observations. This is because the defocus space generating process acts as a high pass filter when we take $\alpha < 0$. Further the inverse diffusion process is inherently unstable and the quantization aggravates the unstability. Practical implementation suggests the use of 16 bit representation of photometric data.

Generation of virtual observations using eqn(7) on local windows may consume quite a bit of computation. This is more due to the fact that a finer sampling of the extended defocus space would lead to a better accuracy in depth estimate. However, to obtain better estimates of the fully focused points efficiently, a hierarchical virtual sampling technique is used wherein, using the algorithm defined earlier a value of α is quickly estimated using coarser discrete steps in the value of α in the range $[\beta, 0]$. Then a further dense sampling is performed in a small neighborhood ϵ around the current best estimate of α, i.e. $\alpha \in [\alpha - \epsilon, \alpha + \epsilon]$

and the estimate of α is refined in a hierarchical manner. This is computationally more efficient and offers a tradeoff between time and accuracy.

6 Results

The algorithm has been tested with real as well as simulated data. In the case of real data, there is a substantial amount of noise in the structure recovered with the depth estimated recovered being irregular. This is mainly because the real world data is in eight bit form and the resultant quantization error plays a major role in the shift in the result. However, the overall structure recovered resembles the general structure of the scene. In a similar way, the corresponding deblurred observation obtained, in general, does not represent the actual pin-hole image, but the result is definitely better focused and more deblurred than the observations given as input to the algorithm. The results obtained with synthetic data can be observed to be better than that obtained with real data due to use of 16 bit representation.

In Fig.3, two images of a ball are taken with varying lens to image plane distances. In the experimental setup the base was at a distance of 117 cms. from the camera. The point on the ball nearest to the camera was at 121.8 cms. while the points lying on the occluding boundary of the ball were at a distance of 132.3 cms. from the camera. The change in the lens-to-image plane distance introduces a small amount of change in magnification. This was taken into account and corrected. The Fig.3.c shows the dense depth map estimated with the darker portions corresponding to nearer distances and the brighter regions corresponding to further distances. The darkest points (grey level 0) refer to the homogeneous region for which the depth cannot be estimated as explained in section 5. The Fig.3.d shows the deblurred image obtained.

The second experimental setup was the "blocks world" where three blocks were arranged at different depths, the nearest one at a distance of 73 cms., another at 82.7 cms. and the farthest block at 96.6 cms. Again images were taken with varying lens to image plane distances to obtain different amont of defocus in different observations. The Fig.4.c shows the dense depth map estimated in this case and the Fig.4.d shows the deblurred image obtained.

The Fig.5 shows a test data where a textured image is synthetically blurred with continuously varying Gaussian functions. The variance of the Gaussian function was increased in a ramp like manner. The variance for the first image E_1 varies from 2 to 7 and for the second image E_2 the variance varies from 1 to 3. The Fig.5.c shows the corresponding dense depth map and the Fig.5.d shows the corresponding deblurred image. The left to right variation in depth is clearly visible.

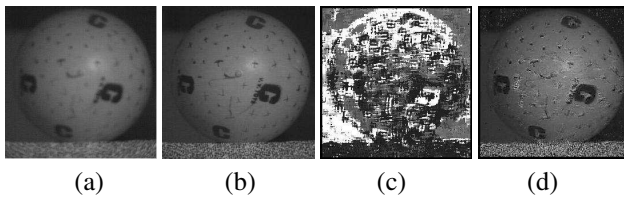The Fig.6 also shows a synthetic test data where a tex-

Figure 3: Ball Image: (a,b) Two observations with the right less blurred, (c,d) recovered structure and the deblurred image.
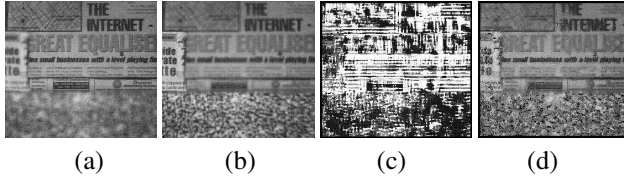

Figure 4: Block World with two observations (a) The furthest block is more in focus, (b)the nearest block is more in focus, (c,d) recovered structure and the deblurred image.
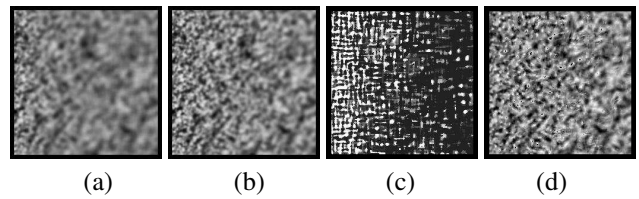

Figure 5: Ramp blur: (a,b) Two synthetically blurred observations, (c) recovered structure, and the (d) deblurred image.
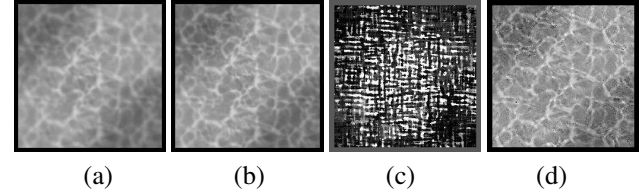

Figure 6: Radial blur: (a,b) Two simulated observations, (c,d) recovered structure and the deblurred image.

tured image is blurred with a continuously varying Gaussian functions. However, here the variance of the Gaussian function was increased in a radially outward manner. The Fig.6.c shows the corresponding dense depth map and the Fig.6.d shows the corresponding deblurred image. Once again the depth variation is quite clear from the plot.

The results appear to be noisy as the linear diffusion process suffers from instability if the defocus space is extended into the process corresponding to inverse diffusion. This method however presents a basis for understanding depth from focus/defocus in the light of the diffusion equation and is thus important in its own merit.

7   Conclusion

For a given scene in the real world, we have defined a defocus space which is a virtual space of all observations based on the properties of a real aperture imaging system. A method for generating the defocus space based on the diffusion equation has been presented. We have also presented an algorithm for recovering structure based on the defocus space. An interesting outcome of this work is that it brings out the equivalence of the depth from focus and depth from defocus modalities for depth estimation. This algorithm has been tested with real as well as synthetic images. A possible extension of the current algorithm is to incorporate a facet based modeling of the depth of the scene while calculating the diffusion coefficient. It is also possible to consider multiple exposures of the scene as is commonly done in the DFF process and this is expected to improve the results. We are also exploring the suitability of regularizing the diffusion process in order to obtain a smooth depth map.

References

[1] S. Chaudhuri and A. N. Rajagopalan. *Depth From Defocus: A Real Aperture Imaging Approach*. Springer Verlag, 1999.

[2] J. J. Koenderink and A. J. van Doorn. Dynamic shape. *Biological Cybernetics*, 53:383–396, 1986.

[3] E. Krotkov. Focusing. *International Journal of Computer Vision*, 1:223–237, 1987.

[4] S. Osher L. I. Rudin and E. Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D*, 60:259–268, 1992.

[5] T. Lindeberg and B. M. ter Haar Romeny. Linear-Scale - Space. In B. M. ter Haar Romeny, editor, *Geometry-Driven Diffusion in Computer Vision*, pages 1–41. Kluwer Academic Publisher,Boston, 1994.

[6] R. Kimmel N. Sochen and A. M. Bruckstein. Diffusions and confusions in signal and image processing. *J. Math. Imaging and Vision*, 14:195–209, 2001.

[7] S. K. Nayar and Y. Nakagawa. Shape From Focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994.

[8] A. Mennucci P. Favaro and S. Soatto. Observing shape from defocused images. *International Journal of Computer Vision*, 52(1):25–43, April 2003.

[9] A. P. Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):523–531, july 1987.

[10] P. Perona and J. Malik. Scale Space and Edge Detection using Anisotropic Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[11] M. Subbarao. Parallel depth recovery by changing camera aperture. In *Proceedings of International Conference on Computer Vision*, pages 149–155, 1988.

[12] M. Subbarao and T. Choi. Accurate Recovery of Three-Dimensional Shape From Image Focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):266–274, march 1995.

[13] D. V. Widder. *The Heat Equation*. Academic Press Inc., London, UK, 1975.

[14] A. P. Witkin. Scale-Space Filtering. In *Proc. of the 4th International Joint Conference on Artificial Intelligence*, pages 1019–1022, Kalsruhe, West Germany, 1983.