Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012

RECOVERING DEPTH FROM IMAGES USING ADAPTIVE DEPTH FROM FOCUS

BING-ZHONG JING, DANIEL S. YEUNG

Machine Learning and Cybernetics Research Center, School of Computer Science and Engineering, South China University of Technology, 510006, Guangzhou, China
E-MAIL: henry_king1986@qq.com

Abstract:

Depth estimation from a sequence of images is a challenging problem in computer vision research. One of the well-known solutions is the depth from focus. However, the drawbacks of this method are the tradeoff between spatial resolution and robustness, and failure in textureless regions. In this paper, a novel approach of depth from focus with multiple images is proposed to improve the two shortcomings. By employing the mean shift segmentation before the step of building Markov random field, the result of segmentation serves as adaptive window for DFF. The edges of the recovered depth map are guaranteed to align with the edges of the original image. After the initial estimation of depth, the hierarchical Markov random field is generated to expand the area to extract depth information according to the structure of the scene. In this way, the experiments show that depth can extract from the textureless regions to some extent.

Keywords:

Depth of field, depth map, depth estimation, mean shift segmentation, Markov Random Field, Depth from Focus

1. Introduction

Three major factors of a photograph, which are aperture, shutter speed, and focus, cannot be altered after an image is being captured in a traditional camera system. A method which allows refocusing (or an extended depth of field) is a potential powerful tool for digital image editing. Once obtaining the depth map, one can deblur the image in order to acquire an all-focus image, or blur the image even more to create certain visual effects [1]. The depth map can also apply to the tasks such as automatic scene segmentation, post-exposure refocusing, and re-rendering of the scene from an alternative viewpoint. By analyzing DOF (depth of field), the coarse depth map of the scene can be recovery [2].

There are many approaches to recover the depth information. Depth from defocus (DFD) or Depth from focus (DFF) are two methods to estimate the 3-d geometry of the scene by exploiting image focus [3].

DFD or DFF based on camera focus and defocus avoid the problem of partial occlusion for the absence of correspondent points matching procedure comparing with methods of stereo and structure from motion [4–5]. DFD captures a sequence of images of a stationary scene with different lens focus settings and attempts to extract depth information from the relative blurriness of these images [2][6-10]. It can recover the depth of the scene given as few as two images. DFF scans the scene by taking a sequence of images with different focus settings and tries to decide a best-focus image for each point. In a general static scene case, DFF is more preferable than DFD since DFF makes a mild assumption about the defocus model and the formation process of the image. The only assumption in DFF is the value of focus measure is minimized at the best focus position.

However, DFF faces the problem of tradeoff between stability and spatial resolution, for focus measures need to be evaluated within a window. A larger window will be more stable while a smaller window gives result with higher spatial resolution [11].

Another problem is that DFF technique is usually designed for highly textured images and fail to generate depth information in the textureless regions. Because the defocus process cannot generate perceivable changes to those textureless surfaces.

In this work, we present a method to perceive the depth from a sequence image captured by a normal camera during its focus process. The mean shift segmentation is applied to the images. Each segment is treated as an adaptive window for focus measure evaluation. Then a hierarchical Markov Random Field is employed to produce more robust depth estimation in textureless regions. Finally, the edges of the depth map is refined by guided image filter.

This paper is organized as follows: Section 2 describes the recent development of DFF and related algorithms. The whole process of depth map extraction is shown in Section

978-1-4673-1487-9/12/$31.00 ©2012 IEEE

3 and tested in Section 4. Finally, we conclude our work in Section 5.

2. Key Concepts

In this section, we introduce recent development of DFF and provide the descriptions of focus measure as well as the mean shift segmentation algorithm.

2.1. Depth From Focus (DFF)

DFF is a 3-dimensional reconstruction method by applying focus measure directly on a set of photos with different focus settings. The best focus setting for a certain region in the scene, which is corresponding with the depth of this region, is decided by the responses of focus measures. The advantage of DFF is its simplicity, which does not need an explicit defocus model.

The implementation of DFF usually changes focus setting while maintains the other parameters of lens. This process can be regarded as employing a lens testing the surface of a 3-dimensional object with different focus settings [12]. DFF requires the camera remains still when capturing images.

The key of successful identifying the peak of responses of focus measures by DFF is enough variation of radiance in the window of focus measure. The performance of DFF is not good when the testing region is textureless or gradient variant. DFF cannot handle the region without texture because there is not enough information to recovery depth. The reason for the failure of the latter case is that no matter which point spread function can produce the same defocus result.

Almost all focus measures analyze an image based on a spatial window with an assumption that all pixels in this window belong to the same focus plane.

Another problem is the light outside the window may contaminate the image in the window due to the defocus process. Because of the effect, the peak of the response the focus measure may be shift. The problem can be eliminated by utilizing a window larger than the blur kernel. However, larger the window is, lower the spatial resolution of the recovered depth map would be.

2.2. Focus Measure

The focus regions of the scene can be detected even without the prior knowledge of the blur kernel by using focus measure. The principle of focus measure is applying a contrast detector such as laplacian of gradient detector to the image in a spatial window.

Several common focus measures are listed as below:

Variance:

$$M_1 = \frac{1}{N^2}\sum_x \sum_y \left(g_\sigma(x,y) - \mu\right)^2 \tag{1}$$

Gradient:

$$M_2 = \sum_x \sum_y \left(g_x^2 - g_y^2\right) \tag{2}$$

Laplacian:

$$M_3 = \sum_x \sum_y \left(g_{xx}^2 - g_{yy}^2\right) \tag{3}$$

The substance of these focus measures are detecting the high frequency of the image.

2.3. Mean Shift Segmentation

Image segmentation is one of the most important low-level vision operations. Mean shift is an unsupervised clustering algorithm which recursively estimates the gradient of the density function to converge the data to the nearest stationary point, which is known as mode.

Mean shift segmentation is based on nonparametric feature space analysis that can avoid such artifacts. Let $\hat{f}(x)$ be the multivariate kernel density estimator [13-14]:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i) \tag{4}$$

where $K_H(x)$ is the kernel with a symmetric positive definite $d \times d$ bandwidth matrix H. The multivariate mean shift vector in the position x is given by

$$m_K(x) = \frac{\sum_{i=1}^{N} x_i K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)} - x \tag{5}$$

This is also called mean shift property. By recursively applying the mean shift property to every point in the feature space, modes, which are the local maxima of the density where $m_K(x) = 0$, and the related data points that define the basin of attraction, can be yielded. The boundaries of the basins define the region of the clusters.

In color image segmentation, the algorithm usually uses a 5-dimenstinoal feature space. A uniform color space like L^*u^*v is usually employed for its metric approximates to Euclidean. The other two dimensions are the coordinates of the points in the image.

3. Description of the Depth Estimation Process

Our proposed method for depth estimation is introduced. The procedure of the pre-process by reverse heat equation, refining the result by mean shift segmentation and hierarchical MRF are described in details

in this section.

3.1. Pre-processing by Optical Flow

The presumption of DFF is that the camera should be stable. However, in practice, people do not always set up a tripod when capturing images. Most compact digital cameras focus on the object according to contrast. This process is similar to DFF. Therefore, in our depth estimation system, we apply DFF to images captured during the focus process, in which the camera is usually hand-held. Even if the camera remains fixed, the changes of distance between sensor and lens during the focus process will cause a slight change in image size. As a result, the vibration and the scaling effect of camera in this process should be compensated by image registration.

Optical flow-based image registration method is applied. The difference between two photos can be described as affine transformation.

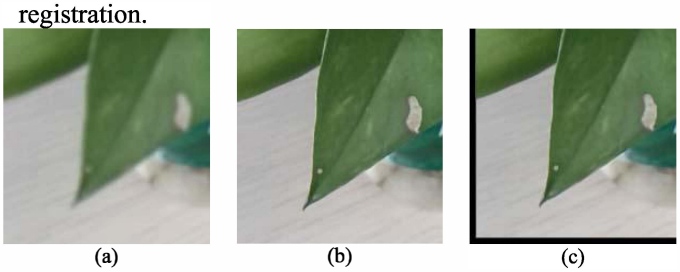
$$F_i(x) = F_j\left(u(x;\theta)\right), \quad i < j \tag{6}$$

$$\begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{7}$$

According to optical flow we have

$$\begin{bmatrix} F^x x_i \\ F^x y_i \\ F^x \\ F^y x_i \\ F^y y_i \\ F^y \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \end{bmatrix} = -F^t + F^x x_i + F^y y_i \tag{8}$$

We can get the affine transformation between two images by solving. (8). Figure 1 shows a result of the image registration.



(a) (b) (c)

Figure 1. Result of image registration (a) an image focused on the background. (b) an image focused on the foreground; (c) the result of image registration with the black edges representing for the shift because of camera vibration.

This example shows a corner of images of a small bonsai, which is capture by a hand-held camera. The left image is captured by focusing on the background while the middle one is focused on the foreground. The right image shows the result of the image registration. The black edges represent the compensation of the scaling effect and vibration.

3.2. Raw Estimation of Depth by Focus Measure

After registration of the image sequence, the depth of the scene is detected by focus measure. The gradient is chosen as the focus measure.

$$ML(x,y) = \frac{\partial^2 I(x,y)}{\partial x^2} + \frac{\partial^2 I(x,y)}{\partial y^2} \tag{9}$$

In Nayar et al. [4] the next step after calculating the focus measure is to sum over the response of focus measure in a spatial window as (10). (11) is used to select the image which has the highest response as the focused one. It is because the estimation result will be unstable if we analyze the focus measure on each pixel.

$$SML(x_0,y_0) = \sum_{(x,y)\in W(x_0,y_0)} ML(x,y) \tag{10}$$

$$D(x,y) = \arg\max_i \left(SML_i(x,y)\right) \tag{11}$$

According to [11], smaller window leads to higher spatial uncertainty and less tolerance of noise while larger window leads to higher uncertainty in frequency domain, more tolerance of noise, but lower spatial resolution. To avoid such a problem, we analyze focus measure in each segment after mean shift segmentation instead of using a fixed window,

$$SML'(x_0,y_0) = \sum_{(x,y)\in S(x_0,y_0)} ML(x,y) \tag{12}$$

$$D'(x,y) = \arg\max_t \left(SML'_t(x,y)\right) \tag{13}$$

where (x,y) locates at segment $S(x,y)$.

DFF is usually applied to the region with rich texture because textureless segments do not have enough depth information. Therefore, the result will be plausible in the region with rich texture while it will be unstable to sum over the focus measure in the textureless segments. As a result, we need to weight the focus measures according their confidences.

$$SML'(x_0,y_0) = \sum_{(x,y)\in S(x_0,y_0)} c(x,y)ML(x,y) \tag{14}$$

$$D'(x,y) = \arg\max_t \left(SML'_t(x,y)\right) \tag{15}$$

$$c(x,y) = 1 - \exp\left(-\frac{w^2(x,y)}{\sigma_c^2}\right) \tag{16}$$

$$w(x,y) = \sum_{(u,v)\in R(x,y)} \left(\nabla_x I(u,v) + \nabla_y I(u,v)\right) g_\sigma(u,v) \tag{17}$$

where ∇ means gradient, g_σ is Gaussian filtering,

$R(x,y)$ represents the neighbors of point (x,y), and σ_c is the variance of w.

These two ways to evaluate the focus measures in a segment will be compared in Section 4

3.3. Hierarchical Markov Random Field

In order to have a better depth estimation result in textureless regions, a hierarchical MRF framework is sued to refine the raw estimation result.

Mean shift segmentation algorithm is applied before using the step of Markov random field. The advantage of applying mean shift segmentation is that instead of looking at neighbors of a pixel, mean shift segmentation searches a larger region controlled by the parameter h_s [15] in the feature space. Moreover, mean shift segmentation has the property of over-segmentation, which can preserve both boundaries and sufficient details of the image structure. Therefore, the result of mean shift segmentation can be applied as adaptive windows for focus measure.

However, ever if a region of pixels is used to decide which label of depth the segment belongs to, there is still a chance that the segment has insufficient information due to the texturelessness. The reason is that the information extracted from a segment, even it is a larger area than a pixel, is the local cue. Global information is required to decide the depth of the textureless areas.

The mean shift segmentation algorithm on different scales of the image is calculated. The patches with identical color and texture are assumed belong to the same object, which means similar depth. The similar regions trend to group in the same segment in the higher scale of mean shift segmentation. By this way, we can infer the global structure of the scene.

The graph cut algorithm [16] is used to solve the Markov random field between two adjacent scales, from the higher one to lower one.

The result of estimate depth $D'(x,y)$ using (13) or $D''(x,y)$ using (14) is an estimated depth map of $I(x,y)$. There are N levels of hierarchy, thus $\hat{D}_i = \{D'_{i,1}(x,y), D'_{i,2}(x,y), \cdots, D'_{i,N}(x,y)\}$ for each segment. The problem can be modeled by Markov random field which consists of a local energy term and a pairwise energy term.

$$E(\overline{D}) = \sum_i E_1(\overline{D}_i) + \sum_{i,j} E_2(\overline{D}_i, \overline{D}_j) \tag{18}$$

For a depth of each segment, the local energy term is

$$E_{1,p}(\overline{D}_i) = c_{t,p}(\overline{D}_i - D'_{i,p})^2 \tag{19}$$

This is because when the segment has higher confidence, it is more likely that the ground true value of depth is closed to the estimated depth and a higher penalty should be given.

The depth values of two consecutive layers are combined:

$$E_1(\overline{D}_i) = \frac{E_{1,u}(\overline{D}_i) + E_{1,u+1}(\overline{D}_i)}{2} \tag{20}$$

The pairwise energy term between two segments need to be lower if these two segments have a considerable disparity to encourage assigning different labels to the two segments, for a considerable disparity would be considered as a sign that these two segments belongs to two distinct objects

$$E_2(\overline{D}_i, \overline{D}_j) = \sum_{j \in S'(i)} a_j \left(\overline{D}_i - d_f(i,j)\right)^2 \tag{21}$$

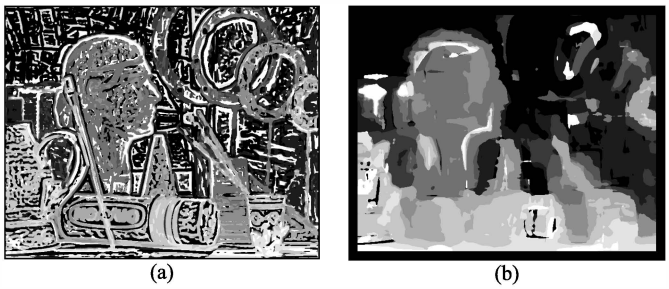
$$d_f(i,j) = 1 - \exp\left(-\frac{(y_i - y_j)^2}{\sigma_{d_f}^2}\right) \tag{22}$$

where y_i is a dimension feature representing segment *i*. This can incorporate the hierarchy information to the graph cut process. After repeating this graph cut algorithm on every two neighboring scales, the 1X scale of depth map can be achieved.

3.4. Edges Refine by Guided Image Filter

Although mean shift segmentation which has the property of over-segmented to protect details at edges, depth discontinuities which have noisy artifacts sometimes still occur due to limitations of the segmentation. Therefore, guided image filter is employed to refine the obtained depth map [17]. In one of our experiments, the images obtained by a camera on an Android mobile phone include a sequence of low-resolution images of focus process and one high-resolution final image. As a result, the depth map estimated from those low-resolution images need to upsample to high-resolution to fit the final image aiming by guided image filter.

4. Results



(a) (b)

Figure 3. An example of different window sizes for DFF

In the first test, we use lens blur filter in Photoshop to simulate the defocus blur on 10 scenes. Each scene includes 10 samples with focal plane from far to nearby. In a common DFF method, the windows for evaluate focus measures are fixed. Figure 3 shows the results for the fixed-window DFF. Figure 3(a) has a smaller window size and higher spatial resolution comparing with Figure 3(b). However, it also has less tolerance to noise than Figure 3(b).

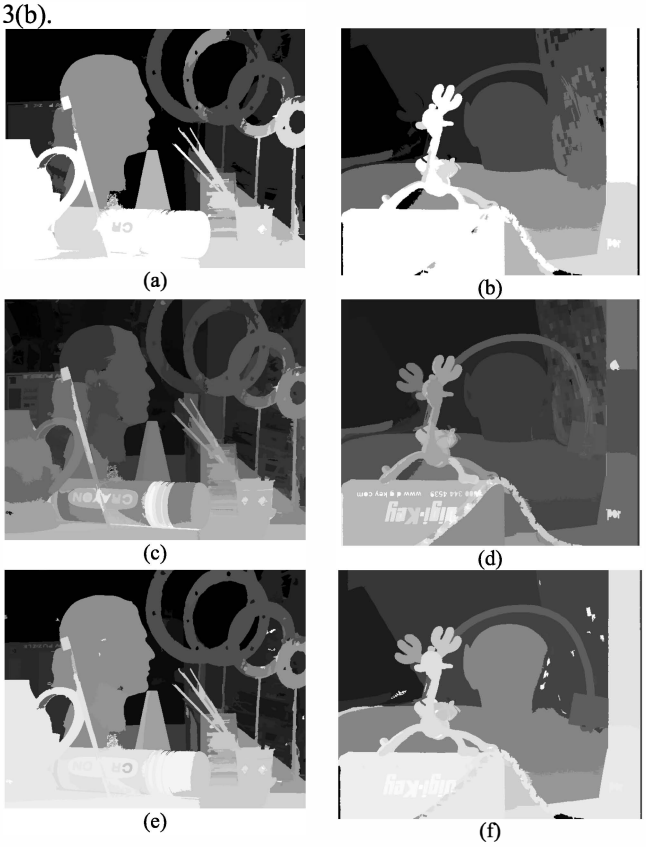


Figure 4. Two test cases. (a), (b) Raw estimation of depth with Eqn. 13; (c), (d) Raw estimation of depth with Eqn. 14; (e), (f) Refine (a), (b) by hierarchical MRF

As we describe in (13) and (14), every pixel in region is either treated evenly or differently with their confidences. Figure 4 shows the different estimated results by these two methods. The result of (13) is better than the one using (14) in the face region and the box area the previous method successfully labels the surface of these objects at the same depth. But there are minor errors need to be fixed, like the ear of the toy deer and the rings in the background of the first image.

Figure 4(e) and (f) shows the results of apply hierarchical MRF. In figure 4 (f), the depth of the deer's ear has been relabled according to their distances in spatial space and feature space. Finally new depth values of two ears are closed which corresponds with the practical situation. After the refining process, the rings in the background also have simliar depth values.

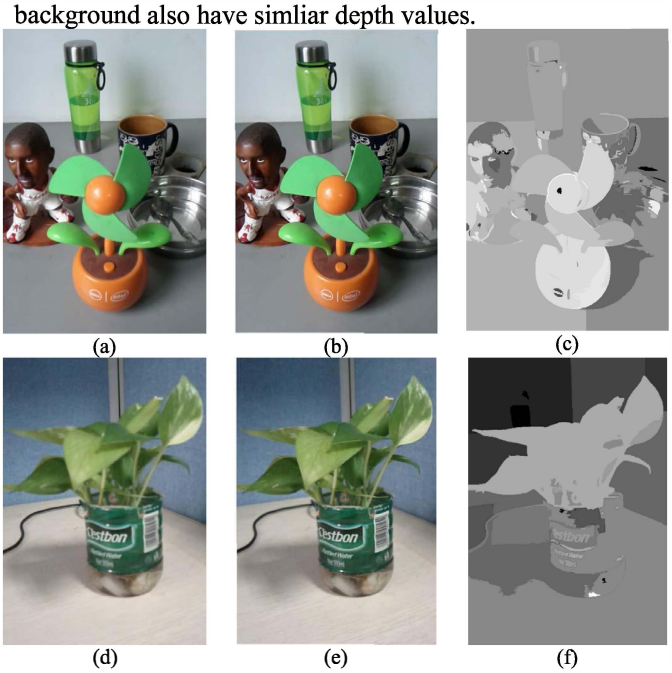


Figure 5. Samples collected by Android mobile phone's camera. (a), (d) Focus on the background; (b), (e) Focus on the foreground; (c), (f) depth estimated by Eqn. 14.

The samples of the second test are collected from an Android mobile phone's camera. We hand-helded the camera and captured all the images of the focus process of each scene. Figure 5 shows two cases of our testing samples. In these real scenes, the result of using the first method defined by (13) is not satisfying while the latter one works much better. Probably in the real cases, the defocus variance is much smaller than the simulated cases and the influence of noise is much larger. Since noise would deteriorate the robustness of focus measure, the latter method with confidence can suppress the noise region and emphasize the valid pixels.

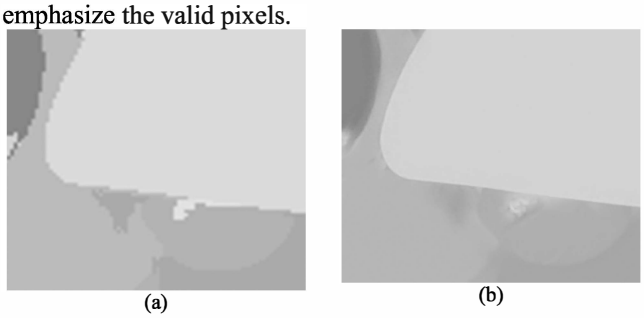


Figure 6. Upsample depth map by guided image filter. (a) Resize the depth map to a high resolution directly; (b) Apply guided image filter to (a)

Since the estimated depth map is smaller than the final image, we employ guided image filter to upsample the estimated depth map, shown in figure 6. After obtaining the high-resolution depth map, large aperutre effect can be simaled by using lens filter or DOF filter in Photoshop as shown in figure 7

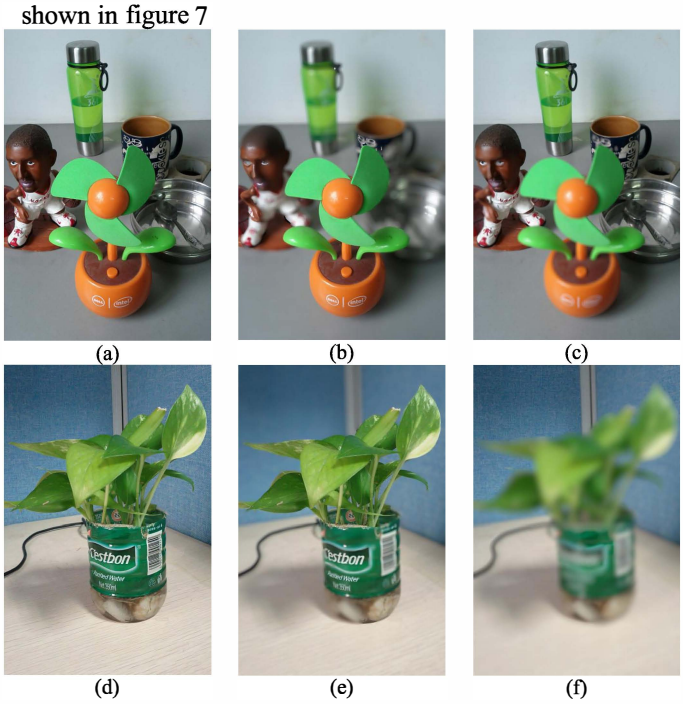


Figure 5. Magnify defocus by Photoshop lens filter. (a), (d) the narrow-DOF images captured by camera; (b), (e) Magnify defocus and focus on the foreground; (c), (f) Magnify defocus and focus on the background.

5. Conclusions

In this paper, we illustrate an adaptive DFF depth estimation method to extract depth map from image sequence captured in a narrow depth of field setting. This method uses the segments produced by mean shift segmentation as windows to analyze the focus measure and employs the proposed hierarchical MRF to infer the depth map. In the real scene cases, after upsampling the depth map by guided image filter, the depth of field of image can be extended. The experimental results show that this depth estimation technique is reliable. In future work, we will incorporate user input and scene detection to improve the accuracy of the estimation results.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61003171 and 61003172), the Fundamental Research Funds for the Central Universities 2011ZM0066 and a Program for New Century Excellent Talents in University (No. NCET-11-0162).

References

[1] POTMESIL, M., AND CHAKRAVARTY, I. 1981. A lens and aperture camera model for synthetic image generation. In Proc. SIGGRAPH, 297-305.

[2] PENTLAND, A. P. 1987. A new sense for depth of field. IEEE Trans. Pattern Anal. Mach. Intell. 9, 4, 523-531.

[3] HASINOFF S W, KUTULAKOS K N. A layer-based restoration framework for variable-aperture photography, F, 2007 [C]. IEEE.

[4] NAYAR, S. K., AND NAKAGAWA, Y. 1994. Shape from focus. IEEE Trans. Pattern Anal. Mach. Intell. 16, 8, 824-831.

[5] ASADA, N., FUJIWARA, H., AND MATSUYAMA, T. 1998. Edge and depth from focus. Int. J. Comput. Vision 26, 2, 153-163.

[6] SUBBARAO, M., AND SURYA, G. 1994. Depth from defocus: A spatial domain approach. Int. J. Comput. Vision 13, 271-294.

[7] GROSSMANN, P. 1987. Depth from focus. Pattern Recognition Letters 5, 1 (Jan.), 63-69.

[8] HASINOFF, S. W., AND UTULAKOS, K. N. 2006. Confocal stereo. In European Conference on Computer Vision, I: 620-634.

[9] FAVARO, P., MENNUCCI, A., AND SOATTO, S. 2003. Observing shape from defocused images. Int. J. Comput. Vision 52, 1, 25-43.

[10] CHAUDHURI, S., AND AJAGOPALAN, A. 1999. Depth from defocus: A real aperture imaging approach. Springer-Verlag, New York.

[11] XIONG Y, SHAFER S A. Depth from focusing and defocusing, F, 1993 [C]. IEEE.

[12] NAIR H N, STEWART C V. Robust focus ranging, F, 1992 [C]. IEEE.

[13] COMANICIU, D., AND MEER, P. 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 5, 603-619.

[14] J.N. Kaftan, A.A. Bell and T. Aach, Mean shift segmentation evaluation of optimization techniques, Proc. Computer Vision Theory & Applications, INSTICC, Madeira, Portugal (2008).

[15] Christopher M. Christoudias, Synergism in Low Level Vision, Proceedings of the 16 th International

Conference on Pattern Recognition (ICPR'02) Volume 4, p.40150, August 11-15, 2002

[16] Yuri Boykov, Olga Veksler, Ramin Zabih, Efficient Approximate Energy Minimization via Graph Cuts. IEEE transactions on PAMI, vol. 20, no. 12, p. 1222-1239, November 2001.

[17] Kaiming He, Jian Sun, and Xiaoou Tang, Guided Image Filtering.The 11th European Conference on Computer Vision (ECCV 2010)