Depth Estimation Using Defocused Stereo Image Pairs

Uma Mudenagudi*

Electronics and Communication Department
B V B College of Engineering and Technology
Hubli-580031, India.

Subhasis Chaudhuri

Department of Electrical Engineering
Indian Institute of Technology-Bombay
Mumbai - 400076, India.
Email-sc@ee.iitb.ernet.in

Abstract

In this paper we propose a new method for estimating depth using a fusion of defocus and stereo, that relaxes the assumption of a pinhole model of the camera. It avoids the correspondence problem of stereo. Main advantage of this algorithm is simultaneous recovery of depth and image restoration. The depth (blur or disparity) in the scene and the intensity process in the focused image are individually modeled as Markov random fields (MRF). It avoids the windowing of data and allows incorporation of multiple observations in the estimation procedure. The accuracy of depth estimation and the quality of the restored image are improved compared to the depth from defocus method, and a dense depth map is estimated without correspondence and interpolation as in the case of stereo.

1 INTRODUCTION

In recent years, an important area of research in computer vision has been the recovery of 3D information about a scene from its 2D images. In the case of human vision, there is the concept of binocular fusion, when stereoscopically presented images appear as a single entity. Julesz[1] showed that random dot stereograms provide a cue for disparity even when an individual image does not provide any high level cue for depth. Pentland [2] reported that the gradient of focus inherent in biological and most optical systems is actually a useful source of depth information. Conventional stereo analysis assumes an ideal pin-hole camera model which offers an infinite depth of field. Any practical camera system is bound to provide depth related blurring in images, which itself is an important cue. Hence, in this paper we fuse stereo and defocus cues to obtain an improved accuracy.

Binocular stereo matching is, in general, ambiguous if the matching is evaluated independently at each point purely by using image properties. All stereo matching algorithms examine the candidate matches by calculating how much support they receive from their local neighborhood. Marr and Poggio[3] proposed a cooperative stereo algorithm based on a multi resolution framework. Barnard and Thompson[4] proposed a feature-based iterative algorithm to solve the correspondence problem. A large number of papers have appeared in the literature on stereo analysis and a review of them can be found in [5].

Let us now look at the literature on depth recovery from defocused images. In [6] Subbarao proposed a more general method compared to that of Pentland [2] in which he removed the constraint of one of the images being formed with a pin-hole aperture. In [7], Xing and Shafer proposed two methods, one is depth from focusing and the other is depth from defocusing. In depth from defocus, they proposed a new camera calibration model, by considering geometric as well as imaging blur. Rajagopalan and Chaudhuri proposed various methods, for example, a block shift-variant blur model[8] that incorporates the interaction of blur among neighboring subregions. Space variant (SV) approaches for depth recovery using a space-frequency representation framework are given in [9],[10]. They have also proposed a method [11] of estimating space variant blur as well as the focused image of the scene from two defocused images. In this method, both the focused image and the blur are modeled as separate MRFs and their MAP estimates are obtained using simulated annealing (SA) [12].

Computationally efficient methods are available in the literature for stereo analysis. Kanade and Okutomi [13] have given a new stereo matching algorithm with an adaptive window, the size of the window is selected by evaluating the local variation of the intensity and the disparity. In [5], a nonlinear diffusion

*Work done at the Department of Electrical Engineering, Indian Institute of Technology-Bombay, Mumbai, India.

is used to estimate the window size. The accuracy of estimates in depth from defocus (DFD) methods is inferior to that of stereo based methods, while in stereo, setting up the correspondence is a difficult task. In this paper we fuse these two methods to estimate the depth information for an improved accuracy. Tsai *et al.* [14] proposed a scheme of integrating stereo and defocus. But they have used rough depth estimates obtained from defocus as a guideline for the stereo matching algorithm. A comparative analysis of DFD and stereo based methods can be found in [15].

As we know in stereo the disparity is directly related to depth. In DFD the blur parameter σ is also directly related to the depth. Hence disparity, a function of σ, is known in terms of lens settings and the base line distance. This information is used to fuse the two methods, thereby getting the advantages of both the methods. In the proposed method, given four images of a scene, *ie,* two defocused stereo pairs of images, we estimate the focused image of the scene and a dense depth (blur or disparity) map using an MAP-MRF approach. The computational problem for the MAP-MRF is solved using simulated annealing.

2 FUSION OF DEFOCUS AND STEREO

In this proposed method we are simultaneously estimating blur (or disparity) and restoring one of the focused image of the scene in the stereo pair (say, the left image). Estimating the other stereo pair is trivial once we know the disparity. As in the most literature, we assume the epipolar line constraint so that the disparity is only in the y-direction. For the given observation model, the right image is given by

$$f_R(x,y) = f_L(x, y + d(x,y)) + w(x,y), \qquad (1)$$

where $d(x,y)$ is the disparity associated with the stereo pair at a point (x,y) and w is the white Gaussian noise. We continue to assume that there is no difference in scene illumination between the left and the right images. The basic structure of the proposed method is given in figure 1. Let us denote by L1 = left image with $\sigma_1(x,y)$ as a blur parameter, L2 = left image with $\sigma_2(x,y)$ as a blur parameter, R1 = stereo pair of L1 with same blur parameter $\sigma_1(x, y+d(x,y))$, R2 = stereo pair of L2 with same blur parameter $\sigma_2(x, y+d(x,y))$. For the DFD camera setup, we also have (see [12] for details)

$$\sigma_1(x,y) = \alpha\sigma_2(x,y) + \beta, \qquad (2)$$

where α and β are known constants that depend on camera settings. The relative blur between the two defocused images is estimated using the intensity information by assuming an appropriate model for the optical transfer function. Usually a Gaussian shaped blur model is assumed. Though the Gaussian blur is of infinite extent, a finite spatial extent approximation ($\pm 3\sigma$ pixels) is assumed for Gaussian blurring windows. We note that the blurring PSF given by $\sigma_i(x,y)$, i=1,2 is space varying and it is directly related to the depth in the scene for a fixed camera setting. The depth (D) is related to the disparity, the baseline distance (b) and the focal length of the camera. If the focal length of the camera is changed, then for the same depth the disparity changes. Let d_m be the disparity and f_m be the focal length for an ideal pin-hole camera associated with the image with blur parameter σ_m, $m = 1, 2$. From the stereo analysis we can write,

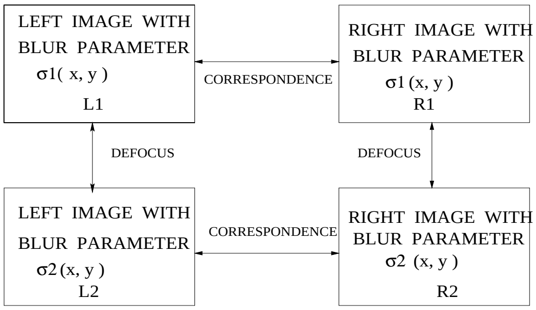


Figure 1: Basic structure of the depth from defocused stereo.

$$d_m = \frac{bf_m}{D}, \qquad m = 1, 2. \qquad (3)$$

Eliminating D we get,

$$d_1 = \frac{f_1}{f_2}\, d_2. \qquad (4)$$

Similarly for a DFD system, the relationship between the blur parameter σ_m, the focal length F_m, the aperture r_m and the lens to image plane distance V_m is given by [12]

$$\sigma_m = \rho r_m V_m \left(\frac{1}{F_m} - \frac{1}{V_m} - \frac{1}{D}\right), \qquad m = 1, 2, \qquad (5)$$

where ρ is a constant parameter related to the resolution of the CCD camera and the subscript m denotes two different observations.

If we now relax the pin-hole camera model for stereo and substitute the value of depth in terms of disparity, we get the disparity in the DFD equation as a function of blur parameter and camera settings, *ie.*

$$d_m = bf_m(\frac{1}{F_m} - \frac{1}{V_m} - \frac{\sigma_m}{\rho r_m V_m}), \quad m = 1,2. \tag{6}$$

If we assume $f_m = V_m$ (since the focal length f_m in a pin-hole model is nothing but V_m in the DFD system, as defined earlier), then the above equation reduces to

$$d_m = b(\frac{V_m}{F_m} - 1 - \frac{\sigma_m}{\rho r_m}), \quad m = 1,2. \tag{7}$$

From the above analysis, once the blur is estimated, the disparity can be determined from the known camera settings and we get a dense depth map without explicitly solving the correspondence problem.

The estimation problem is addressed under the framework of MAP-MRF approach. Computation based on simulated annealing is carried out for simultaneous recovery of depth estimates and the focused image. The utility of MRF lies in its ability to capture local dependencies and its equivalence to the Gibbs random field(GRF). The space variant blur parameter which is related to depth is modeled as an MRF. The local property of MRF leads to an algorithm which can be implemented in a local and parallel manner.

Let S denote the random field corresponding to space variant (SV) blur parameter $S_{ij} = \sigma_1(i,j)$ in the first observation and F_L denote the random field corresponding to the left focused image f_L (intensity process). Assume that S can take P possible levels and F_L can take M possible levels. S is statistically independent to both F_L and the noise field W. The noise field is assumed to be white Gaussian with zero mean and variance σ_w^2. The relation between the focused image and the defocused image is governed by the observation models, for four observed images g_{l_1}, g_{l_2}, g_{r_1} and g_{r_2} with random fields G_{L_1}, G_{L_2}, G_{R_1} and G_{R_2}, respectively

$$g_{l_k} = H_k f_L + w_k, \tag{8}$$

$$g_{r_k} = H_k(d) f_R + {w'}_k \qquad k = 1,2. \tag{9}$$

where g, f and w represent lexicographical ordering of $g(i,j)$, $f(i,j)$ and $w(i,j)$ respectively. H is the blur matrix corresponding to SV blurring function

$$h(i,j;m,n) = \frac{1}{2\pi\sigma_{m,n}^2} \exp\{\frac{-1}{2\sigma_{m,n}^2}[(i-m)^2 + (j-n)^2]\}.$$

$H(d)$ is same as H with the shift due to disparity. Since blur is space variant, H does not possess the nice property of having a block toeplitz structure. The above problem of recovering f_L given four observations is ill posed and may not yield a unique solution, unless additional constraints like a smoothness are added to restrict the solution space. Since S and F_L are modeled as separate MRFs, we can write

$$P(S = s) = \frac{1}{z^s} \exp\{-U^s(s)\}, \tag{10}$$

$$P(F_L = f_L) = \frac{1}{z^f} \exp\{-U^{f_L}(f_L)\}. \tag{11}$$

The terms $U^s(.)$ and $U^{f_L}(.)$ correspond to the energy functions associated with the space-variant blurring process in the left image and the intensity processes in the left image, respectively. Given a realization of S the blurring function $h_1(.)$ is known and hence the matrix H_1 is known. Moreover, $h_2(.)$ is also determined by $\sigma_{ij_2} = \alpha\sigma_{ij_1} + \beta$. Since the disparity is a function of space variant blur $h_1(.)$, $h_2(.)$ for the right pair is calculated. Now, given the four observed images, the *a posteriori* conditional joint probability of S and F_L is given by,

$$P(S = s, F_L = f_L | G_{L_1} = g_{l_1}, ..., G_{R_2} = g_{r_2}) =$$

$$\frac{P(S{=}s, F_L{=}f_L) P(G_{L_1}{=}g_{l_1}, ..|S{=}s, F_L{=}f_L)}{P(G_{L_1} = g_{l_1}, ..., G_{R_2} = g_{r_2})}. \tag{12}$$

Since S and F_L are assumed to be statistically independent, and from Bayes' rule we can write,

$$P(S = s, F_L = f_L | G_{L_1} = g_{l_1}, ..., G_{R_2} = g_{r_2}) =$$

$$\frac{P(S{=}s) P(F_L{=}f_L) P(G_{L_1}{=}g_{l_1}, ..|S{=}s, F_L{=}f_L)}{P(G_{L_1} = g_{l_1}, ..., G_{R_2} = g_{r_2})}. \tag{13}$$

As discussed before, we pose the problem of simultaneous space-variant blur estimation and image restoration as the following MAP problem.

$$\max_{s,f} P(G_{L_1}{=}g_{l_1}, ...|S{=}s, F_L{=}f_L) P(S{=}s) P(F_L = f_L).$$

For fixed observations with an appropriate regularizing term (say, first order smoothness), one can show that the posterior energy function is given by

$$\begin{aligned} U^P(s, f_L) = {} & \frac{||g_{L1} - H_1 f_L||^2}{2\sigma_w^2} + \frac{||g_{L2} - H_2 f_L||^2}{2\sigma_w^2} \\ & + \frac{||g_{R1} - H_1(d) f_R||^2}{2\sigma_w^2} + \frac{||g_{R2} - H_2(d) f_R||^2}{2\sigma_w^2} \\ & + \int [\lambda_s(s_x^2 + s_y^2) + \lambda_f(f_x^2 + f_y^2)] dxdy \\ & + \lambda_{st} \ || \ g_{R1} - g_{L1}(y + d(x,y)) \ ||^2 \\ & + \lambda_{st} \ || \ g_{R2} - g_{L2}(y + d(x,y)) \ ||^2, \end{aligned} \tag{14}$$

where

$$f_R(x,y) = f_L(x, y + d(x,y)),$$

and λ_s, λ_f are the regularization parameters corresponding to the blur and the intensity processes, respectively. Here λ_{st} stands for how well the stereo image pairs are matched in terms of disparity.

From the above analysis computing MAP estimates is equivalent to minimizing the posterior energy function. Smoothness constraints on the estimates of space-variant blur and the intensity processes are encoded in the potential function. In order to preserve the discontinuities in both the blurring process and the focused image of the scene, line fields are also incorporated into the energy function [16]. The horizontal and vertical binary line fields corresponding to the blurring process and intensity process are denoted by l^s_{ij}, v^s_{ij}, $l^{f_L}_{ij}$ and $v^{f_L}_{ij}$, respectively. The *a posteriori* energy function to be minimized is defined including line fields as $U^P(s, f_L, l^s_{ij}, v^s_{ij}, l^{f_L}_{ij}, v^{f_L}_{ij})$, where the smoothness term in equation 14 can be replaced by

$$\sum_{i,j} \lambda_s[(s_{i,j} - s_{i,j-1})^2(1 - v^s_{i,j}) + (s_{i,j+1} - s_{i,j})^2$$
$$(1 - v^s_{i,j+1}) + (s_{i,j} - s_{i-1,j})^2(1 - l^s_{i,j}) + (s_{i+1,j} - s_{i,j})^2$$
$$(1 - l^s_{i+1,j})] + \sum_{i,j} \lambda_f[(f_{Li,j} - f_{Li,j-1})^2(1 - v^{f_L}_{i,j})$$
$$+(f_{Li,j+1} - f_{Li,j})^2(1 - v^{f_L}_{i,j+1}) + (f_{Li,j} - f_{Li-1,j})^2$$
$$(1 - l^{f_L}_{i,j}) + (f_{Li+1,j} - f_{Li,j})^2(1 - l^{f_L}_{i+1,j})]$$
$$+\gamma_s[l^s_{i,j} + l^s_{i+1,j} + v^s_{i,j} + v^s_{i,j+1}]$$
$$+\gamma_f[l^{f_L}_{i,j} + l^{f_L}_{i+1,j} + v^{f_L}_{i,j} + v^{f_L}_{i,j+1}],$$

where γ_s and γ_f are the penalty terms associated with each line field for the blur and the intensity processes, respectively.

The simulated annealing algorithm is used to obtain the MAP estimates of the SV blur parameter and the focused image simultaneously. The temperature variable is introduced in the objective function. Annealing-cum-cooling schedule is carried out at each iteration with linear cooling. Since the random fields associated with the SV blur and the image are assumed to be statistically independent, the values of blur s_{ij} at every point (i,j) and f_{ij} are perturbed independently. Currently the parameters of MRF models are chosen in an adhoc way. The initial estimates of the blur are obtained from Subbarao's window based method[6]. The *a posteriori* energy function is, in general non-convex, and algorithms based on steepest descent are prone to get trapped in local minima. Hence we chose the simulated annealing (SA) algorithm for minimizing the posterior energy function. It is important to note that the locality property of the posterior distribution is what enables us to successfully employ the SA algorithm.

3 RESULTS

In this section, we present the performance of the proposed method in estimating the space variant blur (depth) and restoring the image. Results of experimentation are presented on a simulated random dot pattern, a corridor image and real images of the lab. The number of discrete levels for SV blur was chosen as 64. For the intensity process, 256 levels were used which is the same as the CCD dynamic range. Defocused versions of random dot pattern were first generated such that $\sigma_{i,j2} = 0.5\sigma_{i,j1}$. The estimates of s and f_L are perturbed by an *i.i.d* Gaussian noise with variances σ^2_s and $\sigma^2_{f_L}$, respectively. Figures 2(a-d) show the four defocused stereo pair of images. The window based method of Subbarao is used as the initial estimate for the proposed scheme (size of window 8x8 pixels). Figure 3(c) shows the initial estimates of the blur $\sigma_1(x,y)$. The *rms* value of the error in the initial estimate of the blur is 0.55. The values of various parameters used in SA algorithm were $T_0 = 10.0$, $\lambda_s = 5000.0$, $\lambda_f = 0.005$, $\lambda_{st} = 0.01$, $\gamma_s = 10.0$, $\gamma_f = 15.0$, $\theta_s = 0.4$, $\theta_f = 25.0$, $\sigma_s = 0.1$, $\sigma_f = 6.0$, annealing iterations=200, metropolis iterations=100, where T_0 is the initial temperature, θ_s and θ_f are thresholds for deciding edges in the blur and image, respectively. Here ${\sigma_s}^2$ and ${\sigma_f}^2$ are variances with which new Gibbs samples are generated. The restored image and the estimated SV blur are shown in figure 3(b) and (d), respectively. The value of *rms* error in estimating the blur process is reduced to 0.12 using the proposed technique. From the figure it is seen that the blur is well captured even at the edges. It is important to note that using the proposed method we have been able to perform simultaneous space variant image restoration.

The algorithm is now tested on a corridor image shown in Figure 4(a-d) in which the ceiling has a less spectral content than the floor. From figure 5(a), while restoring the image using defocus alone, the estimates are poor at places of large blur which do not have enough spectral content. Results were improved with the proposed scheme as shown in Figure 5(b), since it fuses stereo also. Figures 5(c) and (d) show the estimates of the depth using only the defocus method and the proposed scheme (darker gray level indicates more depth). Again the estimates are poor where there is a less spectral content. Estimates at the ceiling of the corridor were poor since it is a homogeneous region without any spectral content. The *rms* error in estimating the blur process is reduced from 0.78 to 0.34.

Finally the performance of the proposed scheme was tested on a real image data set. Figures 6(a-d) show the left and the right defocused pairs of images. The restored focused images using the DFD alone and the proposed scheme are shown in Figures 7(a) and (b), respectively. The left defocused pair is used to find the initial depth estimates using the window based method. Figures 7(c) and (d) show the estimates of depth from the DFD alone and that from the proposed method, respectively. From figures it is clear that the proposed scheme gives better estimate of the focused image when image has more blur. The planer nature of the depth variation in the scene is more visible from the result of the proposed method.

4 CONCLUSIONS

We have proposed a new method of fusing the DFD and the stereo based methods to improve the accuracy of the depth estimation. The method uses the advantages of both the DFD and the stereo. The *rms* error in the estimates of space varying blur is reduced compared to the DFD method alone. One can simultaneously restore the image of the scene also. The recovered depth map is dense and no separate interpolation or feature matching is required. The method can be easily extended to multiple observations by adding additional terms in equation 14 appropriately. Currently we are looking at ways to speed up the computation.

References

[1] B.Julesz, "Binocular depth perception without familiarity cues," in *Science*, vol. 145 no.3629, pp. 352–362, July 1964.

[2] A.Pentland, T. Darrell and W.Huang, "A simple real-time range camera," in *Proc IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pp. 256–261, 1989.

[3] D. Marr and T.Poggio, "Cooperative computation of stereo disparity," in *Science 194,1976*, pp. 238–247, 1976.

[4] S. Barnard and W. Thompson, "Disparity analysis of images," in *IEEE Trans. PAMI, Vol. 2,No. 4, July 1980,*, pp. 333–339, 1980.

[5] D. Scharstein and R. Szelislci, "Stereo Matching with Nonlinear diffusion," in *Proc IEEE Intl Conf on Computer Vision and Pattern Recognition*, pp. 343–350, 1996.

[6] M. Subbarao, "Parallel depth recovery by changing camera parameters," in *in Proc.IEEE Intl. Conf. on Computer Vision, Florida, USA*, pp. 149–155, 1988.

[7] Y. Xing and S. A. Shafer, "Depth from focusing and defocusing," in *Proc IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pp. 68–73, 1993.

[8] A. Rajagopalan and S.Chaudhuri, "A block shift-variant blur model for recovering depth from defocused images," in *in Proc.IEEE Intl. Conf. on Image Processing, Washington,D.C ., vol.3 Oct.95*, pp. 636–639, 1995.

[9] A. Rajagopalan and S.Chaudhuri, "Space-variant approaches to recovery of depth from defocused images," in *Computer Vision and Image Understanding , vol-68 no-3,Dec-97*, pp. 309–329, 1997.

[10] A. Rajagopalan and S.Chaudhuri, "A variational approach to recovering depth from defocused images," in *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19,no.10, pp. 1158–1165, Oct 1997.

[11] A. Rajagopalan and S.Chaudhuri, "Optimal recovery of depth from defocused images using an MRF model," in *in Proc.IEEE Intl. Conf. on Computer Vision, Bombay, India*, pp. 555–562, 1998.

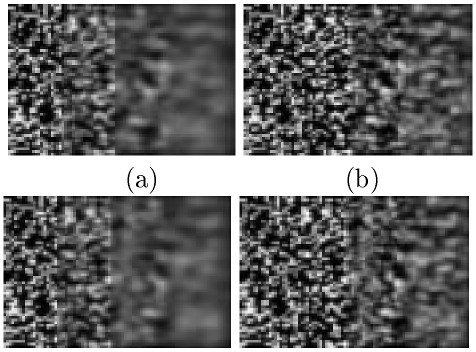
[12] S. Chaudhuri and A. Rajgopalan, *Depth from Defocus: A Real Aperture Imaging Approach.* New York: Springer, 1999.

[13] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," in *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 16, no.9, pp. 920–931, sept 1994.

[14] Y.-P. H. Chun-Jen Tsai, Jit-Tian Leu and C.-H. Chen, "Depth Estimation by the Integration of Depth-from-Defocus and Stereo Vision," in *Institute of Information Science, Academia sinia,Taipei*, pp. 1–28, 1998.

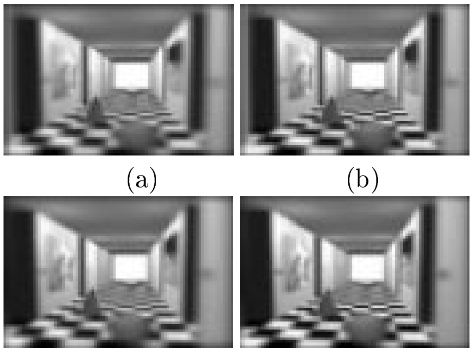
[15] Y. Y. Schechner and N. Kiryati, "Depth from defocus vs. Stereo: How different really are they?," in *Department of Electrical Engineering, Israel Institute of Technology Haifa , Israel*, 1998.

[16] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian distribution of images," in *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, no.6, pp. 721–741, 1984.


(a) (b)

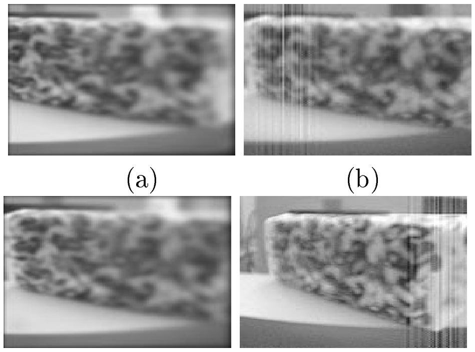
(c) (d)

Figure 2: Left defocused image with (a) blur σ_1, (b) blur σ_2. (c,d) Stereo pair of (a,b), respectively.


(a) (b)

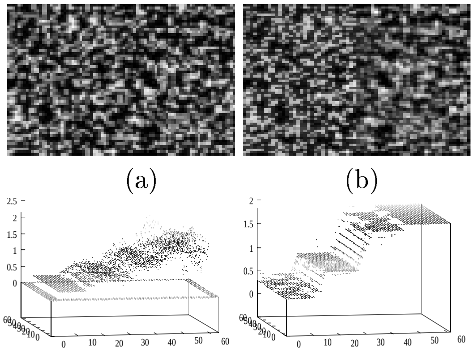
(c) (d)

Figure 4: (a) Left defocused image. (b) Left defocused image with different camera parameters. (c,d) Stereo pair corresponding to (a,b).


(a) (b)

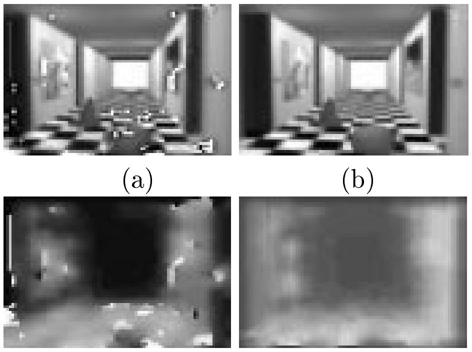
(c) (d)

Figure 6: (a) Left defocused image . (b) Left defocused image with different camera setting. (c) Right defocused image, *ie.* stereo pair of (a). (d) Stereo pair of (b).


(a) (b)

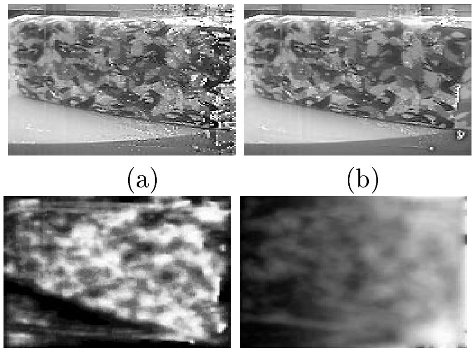
(c) (d)

Figure 3: (a) Original focused image. (b) Reconstructed pin-hole image using proposed method. (c) Initial values of $\sigma_1(x, y)$. (d) Final estimate of $\sigma_1(x, y)$.


(a) (b)

(c) (d)

Figure 5: Reconstructed image for figure 4 using (a) only DFD method, (b) proposed method. Estimated values of σ_1 using (c) only DFD scheme, (d) proposed scheme.


(a) (b)

(c) (d)

Figure 7: Reconstructed image for figure 6 using (a) only DFD method, (b) proposed method. Estimated values of depth using (c) only DFD scheme, (d) proposed scheme.