DEPTH PERCEPTION MODEL EXPLOITING BLURRING CAUSED BY RANDOM SMALL CAMERA MOTIONS

Norio Tagawa, Yuya Iida and Kan Okubo
Graduate School of System Design, Tokyo Metropolitan University, Hino-shi, Tokyo, Japan

Keywords: Depth Perception, Shape from Blurring, Stochastic Resonance, Fixational Eye Movement.

Abstract: The small vibration of the eye ball, which occurs when we fix our gaze on an object, is called "fixational eye movement." It has been reported that this vibration may work not only as a fundamental function to preserve photosensitivity but also as a clue to image analysis, for example contrast enhancement and edge detection. This mechanism can be interpreted as an instance of stochastic resonance, which is inspired by biology, more specifically by neuron dynamics. Moreover, researches for a depth recovery method using camera motions based on an analogy of fixational eye movement are in progress. In this study, using camera motions especially corresponding to the smallest type of fixational eye movement called "tremor." We have constructed the algorithms which are defined as a differential form, i.e. spatio-temporal derivatives of successive two images are analyzed. However, in these methods, observed noise of derivatives causes serious recovering error. Therefore, we newly examine a method in which a lot of images captured with the same camera motions are integrated and the observed local image blurring is analyzed for extracting depth information, and confirm its effectiveness.

1 INTRODUCTION

Camera vibration noise is serious for a hand-held camera and for many vision systems mounted on mobile platforms such as planes, cars or mobile robots, and of course for biological vision systems. The computer vision researchers traditionally considered the camera vibration as a mere nuisance and developed various mechanical stabilizations (Oliver and Quegan, 1998) and filtering techniques (Jazwinski, 1970) to eliminate the jittering caused by the vibration.

In contrast, a new vision device, called the Dynamic Retina (DR), which directly takes advantage of vibrating noise generated by mobile platforms to enhance spatial contrast (Propokopowicz and Cooper, 1995). Furthermore, for edge detection, the Resonant Retina (RR) indicating the DR model with the technique based on stochastic resonance (SR) is proposed (Hongler et al., 2003). SR can be viewed as a noise-induced enhancement of the response of a nonlinear system to a weak input signal, for example bistable devices (Gammaitoni et al., 1998) and threshold detectors (Greenwood et al., 1999), and naturally appears in many neural dynamics processes (Stemmler, 1996).

Although DR and RR offer their massive parallelism and the simplicity of their architecture, by considering especially the enough potential of the camera vibration for depth perception, we have proposed shape recovery methods using the camera motion model imitating fixational eye movements (Tagawa and Alexandrova, 2010), (Tagawa, 2010). These methods are constructed based on a differential form, and the gradient method for "shape from motion" (Horn and Schunk, 1981), (Simoncelli, 1999), (Bruhn and Weickert, 2005) is used fundamentally in order to recover dense depth map with low computational cost compared with the methods based on correlation matching. The fixational eye movement is classified into three types as shown in Fig. 1: microsaccade, drift and tremor. Here, we focus on the tremor, which is the smallest one of the three types, to reduce the linear approximation error of the gradient equation. However, in this case, we cannot get enough information to recover accurate depth from successive two images. Therefore, we have to collect the enough information about depth from other sources. Using a lot of images captured with random small motions of camera, which consists of 3-D rotations imitating fixational eye ball motions (Martinez-Conde et al., 2004), many observations can be used at each pixel, i.e. many gradient equations can be used to recover the each depth value corresponding to the each pixel. It should be noted that since the center of the

Tagawa N., Iida Y. and Okubo K..
DEPTH PERCEPTION MODEL EXPLOITING BLURRING CAUSED BY RANDOM SMALL CAMERA MOTIONS.
DOI: 10.5220/0003817403290334
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 329-334
ISBN: 978-989-8565-03-7

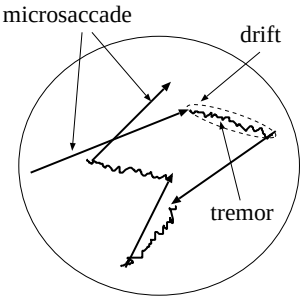

Figure 1: Illustration of fixational eye movement including microsaccade, drift and tremor.

above mentioned 3-D rotations and the lens center differ, the translational motions with respect to the lens center are caused implicitly, and hence, depth information can be observed in these images. Through the simulations using artificial images, if the observation noise is an actual sample of the noise model theoretically defined, the proposed methods work effectively. However, if the size of principal intensity patterns are small as compared with the size of image motions, aliasing occures and hence, the gradient equation becomes useless. This means that the methods in (Tagawa and Alexandrova, 2010) and (Tagawa, 2010) cannot be applied.

In this study, in order to avoid the problem mentioned above, we propose a new scheme based on an integral form using also the analogy of the fixational eye movement. When a lot of images generated by the same way described above are summed up, one blurred image can be obtained. The degree of the blurring is a function of the pixel position, and it also depends on depth value at each pixel. This means that the difference of the degree of burring in image indicates the depth information. By the proposed scheme, at first, using the blurred image detected by summing up all images and the first image with no blurring, spatial distribution of burring in the summed up image is effectively estimated. By modeling the small 3-D rotations of camera as Gaussian random variables, from this burring distribution the depth map can be computed analytically.

2 CAMERA MOTION BLURRING

2.1 Camera Motion Imitating Tremor

As shown in Fig. 2, we use perspective projection as our camera-imaging model. A camera is fixed with an (X,Y,Z) coordinate system, where the viewpoint, i.e., lens center, is at origin O and the optical axis is along the Z-axis. A projection plane, i.e. an image

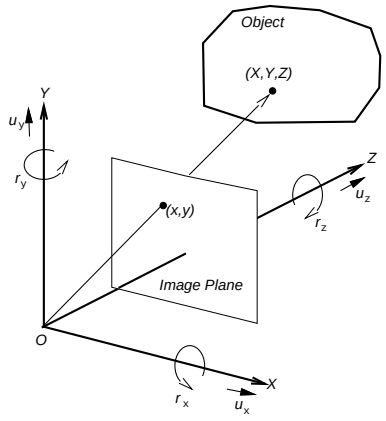

Figure 2: Projection model.

plane, $Z = 1$ can be used without any loss of generality, which means that focal length equals 1. A space point (X,Y,Z) on an object is projected to an image point (x,y).

We introduce a motion model representing fixational eye movement. We can set a camera's rotation center at the back of lens center with Z_0, which is constant and known, along optical axis. Rotations around all axes parallel with X, Y and Z axis, respectively, are considered as a rotation of eye ball. We represent this rotational vector as $r = [r_x, r_y, r_z]^\top$, and it can be used also for the representation of the rotational vector at origin O shown in Fig. 2. On the other hand, the translational vector $u = [u_x, u_y, u_z]^\top$ in Fig. 2 is caused by the above eye ball's rotation, and is formulated as follows:

$$u = r \times \begin{bmatrix} 0 \\ 0 \\ Z_0 \end{bmatrix} = Z_0 \begin{bmatrix} r_y \\ -r_x \\ 0 \end{bmatrix}. \qquad (1)$$

From this equation, it can be known that r_z causes no translation. Therefore, we set $r_z = 0$ and redefine $r \equiv [r_x, r_y]^\top$ as a rotation vector of eyeball. Using Eq. 1 and the inverse depth $d(x,y) = 1/Z(x,y)$, image motion called "optical flow" $v = [v_x, v_y]^\top$ is given as follows:

$$v_x = xyr_x - (1+x^2)r_y - Z_0 r_y d, \qquad (2)$$

$$v_y = (1+y^2)r_x - xyr_y + Z_0 r_x d. \qquad (3)$$

In the above equations, d is an unknown variable at each pixel, and u and r are unknown common parameters for the whole image. This camera model is easy of control, since the degree of freedom for motion is low. Additionally, absolute depth values can be determined by this model with known value Z_0.

We use M as the number of frames used for depth recovery, and in this study, $\{r^{(j)}\}_{j=1,\cdots,M}$ is treated as

a stochastic variable. We ignore the temporal correlation of tremor which is needed to form drift component, and we assume that $r^{(j)}$ is a 2-dimensional Gaussian random variable with a mean 0 and a variance-covariance matrix $\sigma_r^2 I$, where I indicates a 2×2 unit matrix.

$$p(r^{(j)}|\sigma_r^2) = \frac{1}{(\sqrt{2\pi}\sigma_r)^2} \exp\left\{ -\frac{r^{(j)\top} r^{(j)}}{2\sigma_r^2} \right\}, \quad (4)$$

where σ_r^2 is assumed to be known.

From Eqs. 2 and 3, and the probabilistic characteristics of $r^{(j)}$, v is also a 2-dimensional Gaussian random variable with $\mathrm{E}[v] = 0$ and the variance-covariance matrix of

$$V[v] = \begin{bmatrix} x^2y^2 + (1+x^2+Z_0 d)^2 & 2xy(1+\frac{x^2+y^2}{2}+Z_0 d) \\ 2xy(1+\frac{x^2+y^2}{2}+Z_0 d) & x^2y^2 + (1+y^2+Z_0 d)^2 \end{bmatrix} \sigma_r^2. \quad (5)$$

If we can know the variance-covariance matrices depending on image position, depth map can be calculated.

2.2 Image Blurring Related to Depth

There are some schemes to obtain the variance-covariance matrix of optical flow defined by Eq. 5 locally at each image position from multiple images observed through random camera rotations imitating tremor. The most simple and natural way is statistically computing the matrix as an arithmetic average of quadratic value of optical flows, which is firstly detected from images. However, in this study, we suppose that intensity patterns are fine with respect to a temporal sampling rate, and hence optical flow is hard to be detected accurately. Therefore, we employ an integral formed scheme, in which the variance-covariance matrices are computed as a distribution of local image blurring.

We define an averaged image $f_{ave}(x)$ as an arithmetic average of observed M images $\{f_j(x)\}_{j=1,\cdots,M}$ with fixational eye movements. If M is asymptotically large, the following equation holds using locally defined a 2-dimensional Gaussian point spread functions $g_x(\cdot)$ and an original image $f_0(x)$.

$$f_{ave}(x) = \int_{x' \in \mathcal{R}} g_x(x') f_0(x - x') dx', \quad (6)$$

where x indicates the image position, $\mathcal{R}$ is a local region around x, and $g_x(\cdot)$ has a vairance-covaiance matrix in Eq. 5. Additionally, $\int g_x(x')dx' = 1$ is satisfied.

As the above discussion, we model $f_{ave}(x)$ as an image blurred by fixational eye movements. The distribution of blurring degree in $f_{ave}(x)$ represents the spatial distribution of depth.

3 DEPTH PERCEPTION

3.1 Blurring Detection Algorithm

We detect the blurring distribution in an image domain not in a frequency domain. The processing schemes can be classified into an one-step scheme and a multi-step scheme. In the one-step scheme, the unknown value set $\{d_i\}_{i=1,\cdots,N}$ (N indicates the number of pixels in image) is determined with keeping the whole constraints for them. At all pixels in image, the point spread functions $\{g_x(\cdot)\}$, each of which has a Gaussian form and has the variance-covariance matrix formulated by Eq. 5, have to be determined simultaneously, and as a result, $\{d_i\}$ is optimally obtained.

On the other hand, in the multi-step scheme, firstly at the each pixel, the variance-covariance matrix of the Gaussian distribution is detected with no use of the constraint in Eq. 5. After that, $\{d_i\}$ is determined from the variance-covariance matrices using the constraint in Eq. 5. In this study, in order to confirm the possibility of our integral formed scheme, we employ the latter scheme. Additionally, the Gaussian constraints are relaxed, and the variance-covariance matrix is estimated as simple statistics. In the following, the concrete algorithm is explained.

We use the original image, i.e. the first image $f_0(x)$, and the arithmetic average $f_{ave}(x)$ to determine the image blurring, and the each $f_j(x)$ $(j \neq 0)$ is not used explicitly to save capacity of memory. At first, the Gaussian property is ignored and hence, $w_x(\cdot)$ is used as a point spread function instead of $g_x(\cdot)$. The local support of $w_x(x)$ is defined as a square discrete region with $P \times P$ pixels, and using a dictionary order, P^2-dimensional vector w_i is introduced as a discrete representation of $w_x(\cdot)$, where "i" indicates a pixel index. Additionally, discrete representations of local image intensity of $f_0(x)$ and $f_{ave}(x)$ are defined as f_0^i and f_{ave}^i, which are also P^2-dimensional vectors consist of local intensity values around the pixel i. Using f_0^i, $P^2 \times P^2$ matrix F^i is defined as follows:

$$F^i = \begin{bmatrix} f_0^{i+1} & f_0^{i+2} & \cdots & f_0^{i+P^2} \end{bmatrix}. \quad (7)$$

By ignoring the constraint generally holding for the components of the point spread function $\{w_{i(k)}\}_k$ of blurring, $\sum_k w_{i(k)} = 1$, an objective function for each pixel i can be defined as follows:

$$J_i(w_i) = \left(F^{i\top} w_i - f_{ave}^i \right)^\top \left(F^{i\top} w_i - f_{ave}^i \right). \quad (8)$$

By differentiating $J_i(w_i)$ with respect to w_i, the following solution can be derived.

$$\hat{w}_i = \left(F^i F^{i\top} \right)^{-1} F^i f_{ave}^i. \quad (9)$$

Subsequently, the components of the variance-covariance matrix V_i theoretically corresponding to the variance-covariance matrix of the optical flow defined in Eq. 5 have to be estimated. The each component can be simply estimated as follows:

$$\hat{V}_{i(1,1)} = \sum_{k=1}^{P^2} x(k)^2 \hat{w}_{i(k)}, \tag{10}$$

$$\hat{V}_{i(1,2)} = V_{i(2,1)} = \sum_{k=1}^{P^2} x(k)y(k)\hat{w}_{i(k)}, \tag{11}$$

$$\hat{V}_{i(2,2)} = \sum_{k=1}^{P^2} y(k)^2 \hat{w}_{i(k)}, \tag{12}$$

where $(x(k), y(k))$ means the 2-dimensional coordinate values corresponding to k with the center of the local support as $(0,0)$.

3.2 Depth Perception Algorithm

From Eq. 5 and the estimates computed by Eqs. 10, 11 and 12, equations with respect to the each d_i can be derived as follows:

$$1 + x_i^2 + Z_0 d_i = \sqrt{\frac{\hat{V}_{i(1,1)}}{\sigma_r^2} - x_i^2 y_i^2} \equiv \alpha_i, \tag{13}$$

$$1 + \frac{x_i^2 + y_i^2}{2} + Z_0 d_i = \frac{\hat{V}_{i(1,2)}}{2x_i y_i \sigma_r^2} \equiv \beta_i, \tag{14}$$

$$1 + y_i^2 + Z_0 d_i = \sqrt{\frac{\hat{V}_{i(2,2)}}{\sigma_r^2} - x_i^2 y_i^2} \equiv \gamma_i. \tag{15}$$

Using the mean square criterion, estimate of d_i is determined as

$$\hat{d_i} = \frac{1}{Z_0}\Big(w_\alpha \alpha_i + w_\beta \beta_i + w_\gamma \gamma_i - (w_\alpha + \frac{w_\beta}{2})x_i^2 - (w_\gamma + \frac{w_\beta}{2})y_i^2 - 1\Big), \tag{16}$$

where w_α, w_β and w_γ are the weight respectively corresponding to the each of Eqs. 13, 14 and 15 and $w_\alpha + w_\beta + w_\gamma = 1$ holds. Especially, if $w_\alpha = w_\beta = w_\gamma = 1/3$, Eq. 16 becomes

$$\hat{d_i} = \frac{1}{Z_0}\left(\frac{\alpha_i + \beta_i + \gamma_i}{3} - \frac{x_i^2 + y_i^2}{2} - 1\right). \tag{17}$$

By expanding the right-hand side of Eq. 13 as the Taylor series and extracting the first order term, error component can be formulated as $\delta V_{i(1,1)}/(2\sqrt{V_{i(1,1)}\sigma_r^2 - x_i^2 y_i^2 \sigma_r^4})$, in which $\delta V_{i(1,1)}$ is the detection error in Eq. 10. In the same way, error in Eq. 14 is $\delta V_{i(1,2)}/(2x_i y_i \sigma_r^2)$, and error in Eq. 15 is $\delta V_{i(2,2)}/(2\sqrt{V_{i(2,2)}\sigma_r^2 - x_i^2 y_i^2 \sigma_r^4})$. If it is assumed that $\delta V_{i(1,1)}$, $\delta V_{i(1,2)}$ and $\delta V_{i(2,2)}$ are the Gaussian random variables with the same variance, the following weight can be used to determine d_i as the maximum likelihood estimator.

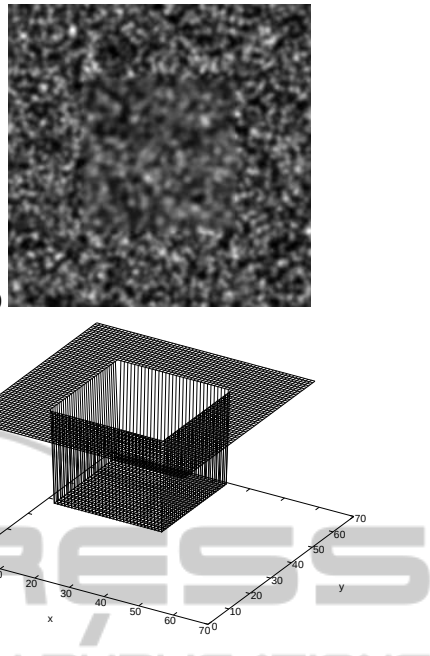

Figure 3: Example of the data used in the experiments: (a) artificial image; (b) true depth map.

$$w_\alpha = \frac{V_{i(1,1)} - x_i^2 y_i^2 \sigma_r^2}{V_{i(1,1)} + V_{i(2,2)} - x_i^2 y_i^2 \sigma_r^2}, \tag{18}$$

$$w_\beta = \frac{x_i^2 y_i^2 \sigma_r^2}{V_{i(1,1)} + V_{i(2,2)} - x_i^2 y_i^2 \sigma_r^2}, \tag{19}$$

$$w_\gamma = \frac{V_{i(2,2)} - x_i^2 y_i^2 \sigma_r^2}{V_{i(1,1)} + V_{i(2,2)} - x_i^2 y_i^2 \sigma_r^2}. \tag{20}$$

4 NUMERICAL EVALUATIONS

To confirm the feasibility of the proposed scheme, we conducted numerical evaluations using artificial images. Figure 3(a) shows the original image generated by a computer graphics technique using the depth map shown in Fig. 3(b). The image size assumed in these evaluations is 256×256 pixels, which corresponds to $-0.5 \leq x, y \leq 0.5$ measured using the focal length as a unit. In Fig. 3(b), the vertical axis indicates the depth Z using the focal length as a unit, and the horizontal axes mean x and y in the image, which is marked every four pixels.

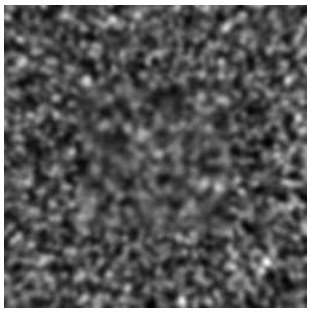

Figure 4: Averaged image with 100 images.

In the evaluations, we generated the successive images from the original image shown in Fig. 3(a) by randomly sampling $r^{(j)}$ as a Gaussian random variable. By varying the number of the images used for averaging and the deviation of $r^{(j)}$, the depth recovery error was evaluated. Figure 4 shows the averaged image $f_{ave}(x)$ with 100 images. The value of P by which the local region size for estimating $V[v]$ is defined is adjusted according to the maximum value of the theoretical $V[v]$ evaluated by Eq. 5, i.e. P is set as $P = \sqrt{\max V[v]} \times 6$. The evaluated characteristics of the recovery error are shown in Figs. 5 and 6. Examples of the depth recovery results are shown in Figs. 7 and 8. In these evaluations, Eqs. 16 with the weights defined by Eqs. 18, 19 and 20 are employed. These weights are defined using the true values of V_i, which can not be known in the actual situation, hence we use the estimated values computed by Eqs. 10, 11 and 12 instead of the true values. Note that Eq. 17 is very poor for a good recovery in this study.

From Figs. 7 and 8, we can confirm that the outline of the recovered depth resembles the theoretical one, but there are a lot of noisy patterns. By increasing the number of images summed up, the depth recovery error becomes small, and hence, the noisy patterns in the estimated $V[v]$ become small a little. On the other hand, when the motion size is too large, the recovery error can not become smaller. This means that using the large motion, the discontinuous regions of the shape may be recovered as the hardly smooth one.

5 CONCLUSIONS

In this study, we propose a new scheme to recover a depth map using the camera model imitating fixational eye movements, especially tremor component. Our scheme is based on the integral form and the image blurring is mainly used to recover depth, although we have proposed some differential-formed methods. We explained the theoretical principle of our scheme and proposed the simple method to estimate image

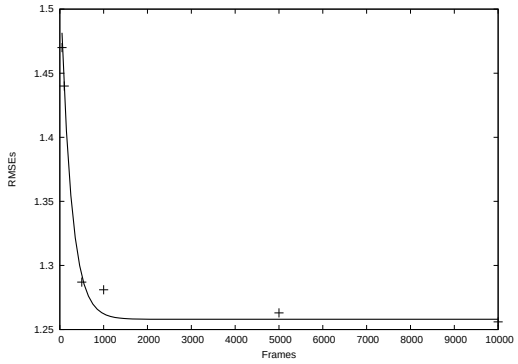

Figure 5: Characteristics of depth recovery error with respect to the number of images.

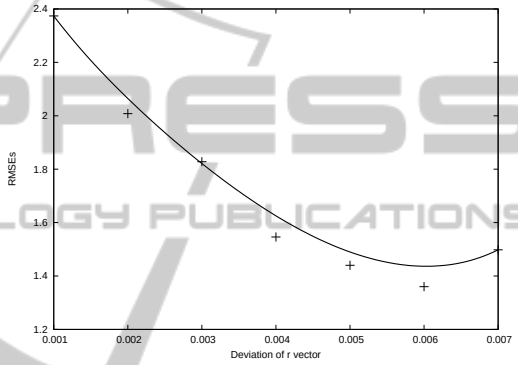

Figure 6: Characteristics of depth recovery error with respect to the deviation of $r^{(j)}$.

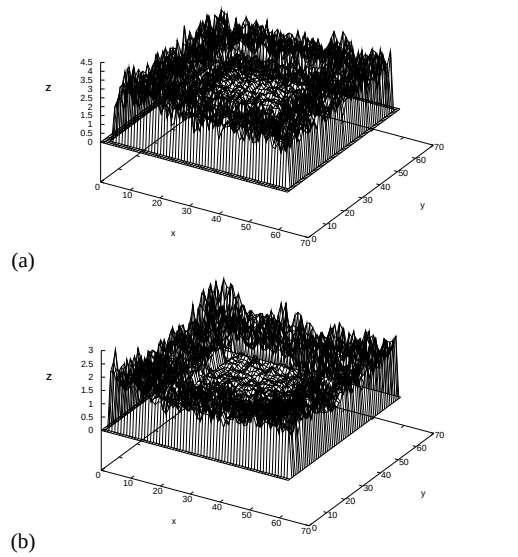

Figure 7: Depth recovery results obtained by varying σ_r with $M = 100$: (a) $\sigma_r = 0.001$ and $P = 3$; (b) $\sigma_r = 0.003$ and $P = 5$.

blurring using the original image and the averaged image. In this method, by simplifying the problem, we ignore the constraints for the blurring of this prob-

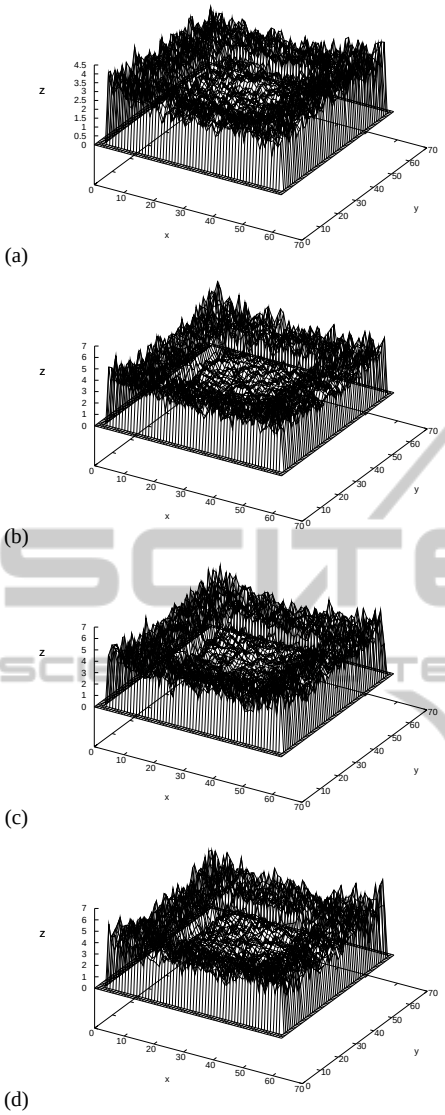

Figure 8: Depth recovery results obtained by varying the number of images with $\sigma_r = 0.005$ [rad./frame] and $P = 9$: (a) $M = 100$; (b) $M = 500$; (c) $M = 1000$; (d) $M = 10000$.

lem, which should be used to recover accurate depth map, and hence, we tried to confirm only the feasibility of the proposed integral-formed scheme. From the results of the numerical evaluations using artificial images, we can know that the proposed scheme can get the depth information. In future, we have to construct the optimal detection method of image blurring caused by the fixational eye movements.

REFERENCES

Bruhn, A. and Weickert, J. (2005). Locas/kanade meets horn/schunk: combining local and global optic flow methods. *Int. J. Comput. Vision*, 61(3):211–231.

Gammaitoni, L., Hanggi, P., Jung, P., and Marchesoni, F. (1998). Stochastic resonance. *Rev. Modern Physics*, 70(1):223–252.

Greenwood, P., Ward, L., and Wefelmeyer, W. (1999). Statistical analysis of stochastic resonance in a simple setting. *Physical Rev. E*, 60:4687–4696.

Hongler, M.-O., de Meneses, Y. L., Beyeler, A., and Jacot, J. (2003). The resonant retina: exploiting vibration noise to optimally detect edges in an image. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(9):1051–1062.

Horn, B. K. P. and Schunk, B. (1981). Determining optical flow. *Artif. Intell.*, 17:185–203.

Jazwinski, A. (1970). *Stochastic processes and filtering theory*. Academic Press.

Martinez-Conde, S., Macknik, S. L., and Hubel, D. (2004). The role of fixational eye movements in visual perception. *Nature Reviews*, 5:229–240.

Oliver, C. and Quegan, S. (1998). *Understanding synthetic aperture radar images*. Artech House, London.

Propokopowicz, P. and Cooper, P. (1995). The dynamic retina. *Int'l J. Computer Vision.*, 16:191–204.

Simoncelli, E. P. (1999). Bayesian multi-scale differential optical flow. In *Handbook of Computer Vision and Applications*, pages 397–422. Academic Press.

Stemmler, M. (1996). A single spike suffices: the simplest form of stochastic resonance in model neuron. *Network: Computations in Neural Systems*, 61(7):687–716.

Tagawa, N. (2010). Depth perception model based on fixational eye movements using byesian statistical inference. In *proc. ICPR2010*, pages 1662–1665.

Tagawa, N. and Alexandrova, T. (2010). Computational model of depth perception based on fixational eye movements. In *proc. VISAPP2010*, pages 328–333.