COMPUTATIONAL MODEL OF DEPTH PERCEPTION BASED ON FIXATIONAL EYE MOVEMENTS

Norio Tagawa and Todorka Alexandrova
Faculty of System Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo, Japan

Keywords: Fixational eye movement, Depth perception, Structure from motion, Bayesian estimation, EM algorithm.

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 function works also as a clue to monocular depth perception. Moreover, researches for a depth recovery method using camera motions based on an analogy of fixational eye movement are in progress. We suppose that depth perception with fixational eye movement is firstly carried out, and subsequently such depth information is supplementary used for binocular stereopsis. Especially in this study, using camera motions corresponding to the smallest type of fixational eye movement called “tremor,” we construct depth perception algorithm which models camera motion as a irregular perturbation, and confirm its effectiveness.

1 INTRODUCTION

Structure from motion is typical for monocular depth perception, and in this case an autonomous motion of human is usually assumed. On the other hand, it is well known that a fixational eye movement, which means an irregular involuntary motion of eyeball, arises when human gazes fixed targets (Martinez-Conde et al., 2004). Since human’s retina can keep sensitivity of receiving by finely vibrating images of targets on a retina, fixational eye movement is the firstly required function to watch something. The human vision system corrects such vibration unconsciously, and recognizes static images. It has been reported that the fixational eye movement plays as a clue for depth perception, regardless of the unconsciousness of image motion caused by it in the retina, and an actual vision system based on a fixational eye movement has been proposed (Ando et al., 2002). This can suggest possibility that unconscious depth perception is performed through a fixational eye movement and the result is inputted into the binocular stereopsis system with the brightness perception and the color perception by binocular system as primitive sources.

A lot of notable results in the study for structure from motion (SFM) have been reported. Although there are various computational principles for SFM, when computatinally efficient and dense depth recovery is considered to be important, the gradient method is effective (Horn and Schunk, 1981), (Simoncelli, 1999), (Bruhn and Weickert, 2005). For the gradient method, it has to be noted that there should be an adequate motion size for each image region in order to recover accurate depth. Since the gradient equation can completely hold when image motion is infinitesimal, the equation error can not be ingored for highly large motion. Inversely for small motion, the motion information is hidden in observation errors of spatio-temporal differentials of brightness, and hence accurate depth can not be recovered. Therefore, it is naturally required to adjust frame rate adaptively in order to make motion size suitable. We have proposed a method with no necessity of variable frame rate, which is based on multi-resolution decomposition of images, but high computational cost is needed (Tagawa et al., 2008). We pay attention to the small motion so as to avoid equation error in the gradient method. To solve the above mentioned S/N problem caused for small motion, we should obtain many observations and use them collectively. For such strategy, motion direction and motion size have to take various values, in order to improve the accuracy independently of the image texture.

Tagawa N. and Alexandrova T. (2010).
COMPUTATIONAL MODEL OF DEPTH PERCEPTION BASED ON FIXATIONAL EYE MOVEMENTS.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 328-333
DOI: 10.5220/0002829203280333

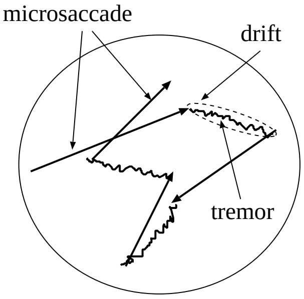


Figure 1: Illustration of fixational eye movement including microsaccade, drift and tremor.

From the above discussions, in this study, we examine a depth perception model based on fixational eye movements. 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, and construct a computation algorithm using analogy of tremor to confirm the effectiveness of the perception model with tremor. Since the fixational eye movement is an involuntary motion, it is realistically hard to know all of the eye movements before depth recovery, and thus we treat them as stochastic variables. This problem can be realized in the framework of the Bayesian inference, and a stable algorithm is expected to be constructed using the EM algorithm (Dempster et al., 1977).

2 PERCEPTION MODEL WITH FIXATIONAL EYE MOVEMENT

As a background of this study, we are examining a two-step perception model in which monocular depth perception based on fixational eye movement is used for binocular stereopsis. Binocular stereopsis plays an essential role in the depth perception of a human vison system (Lazaros et al., 2008), but occulusions often occur in it. By this two-step processing, this occulusion problem is expected to be solved. In this study, we propose mainly a model for the first step perception constructed additionally with the following two-step perception

1. perception in the period of drift and tremor;
2. perception in the period of microsaccade.

In the former, depth perception corresponding to the whole period of one drift, instead of that corresponding to each tremor period, is assumed to be caused by multiple fine movements of tremor over one period of drift. Therefore, recognized depth value has only the temporal resolution equivalent to the period of one drift, and has only the spatial resolution equivalent to the distance of movement of one drift. However, because of treating small movements, the gradient method explained in the next section can be used, which needs no search process and hence, is cost effective. It should be noted that, by adopting drift as an unit of perception, variety of brightness patterns in a neigboring region can be effectively used, and as a result accurate perception of depth can be realized.

In the latter, using the depth value obtained by the former step with low resolution and eye movement corresponding drift, image displacement before and after microsaccade is detected by search process and depth value is recognized. Since the results of the former step can be used, the size of the local region where the brightness pattern is used to search and the range of searching area can be appropriately determined. Additionally, because microsaccade indicates fast movement, by the latter step, depth perception with high spatio-temporal resolution can be done through small computation.

As a first report of our monocular perception model, we construct an algorithm for the first step and confirm its efficiency. To model completely the first step, we have to integrate drift component into the algorithm, but in this study, we focus only on tremor. Hence, we ignore the temporal correlation of tremor which is needed to form drift component, and we assume that each small movements are independent of each other.

3 GRADIENT METHOD USING FIXATIONAL EYE MOVEMENT

3.1 Motion Model and Optical Flow

As shown in Fig. 2, we use perspective projection as our camera-imaging model. The 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. The projection plane, i.e. image plane, $Z = 1$ can be used without any loss of generality, which means that the focal length equals 1. A space point (X,Y,Z) on the object is projected to the image point (x,y). The camera moves with translational and rotational vectors $u = [u_x,u_y,u_z]^\top$ and $r = [r_x,r_y,r_z]^\top$.

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 along optical axis. In this study, we pick out tremor from three

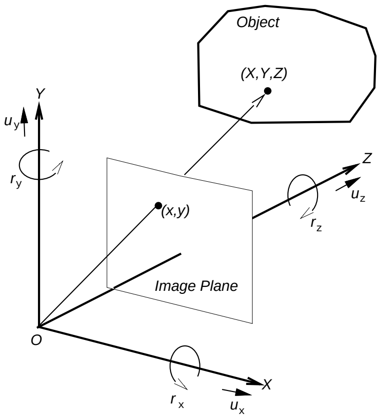


Figure 2: Assumed projection model.

types of fixational eye movement, and hence consider all rotations around all axes parallel with X, Y and Z axis, respectively, as a rotation of eye ball. We represent this rotaion 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 in Fig. 2 is caused by the above eye ball's rotation, and is formulated as follows:

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

Using this representation of u and the inverse depth $d(x,y) = 1/Z(x,y)$, the optical flow $v = [v_x, v_y]^\top$ is given as follows:

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

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

In the above equtions, d is an unknown variable at each pixel, and u and r are unknown common parameters for the whole image.

3.2 Gradient Equation for Rigid Motion

The gradient equation is the first approximation of the assumption that image brightness is invariable before and after the relative 3-D motion between a camera and an object. At each pixel (x,y), the gradient equation is formulated with the partial differentials f_x, f_y and f_t of the image brightness $f(x,y,t)$ and the optical flow as follows:

$$f_t = -f_x v_x - f_y v_y, \qquad (4)$$

where t denotes time. By substituting Eqs. 2 and 3 into Eq. 4, the gradient equation representing a rigid motion constraint can be derived explicitly

$$\begin{aligned} f_t &= -(f_x v_x^r + f_y v_y^r) - (-f_x r_y + f_y r_x) Z_0 d \\ &\equiv -f^r - f^u d. \end{aligned} \qquad (5)$$

In Eq. 5, f_x, f_y and f_t are observations and contain observation noise. Additionally, equation error, i.e. error caused by the first approximation in Eq. 4 generally exists.

3.3 Definition of Probabilistic Model

We use M as the number of pairs of two successive frames and N as the number of pixels. In our study, $\{f_t^{(i,j)}\}_{i=1,\cdots,N; j=1,\cdots,M}$ and $\{r^{(j)}\}_{j=1,\cdots,M}$ are treated as stochastic variables, and $\{d^{(i)}\}_{i=1,\cdots,N}$ corresponding to the inverse depth at each pixel is treated as a definite variable and is recovered independently at each pixel. However, since multiple frames vibrated by irregular rotation $\{r^{(j)}\}$ are used for processing and no tracking procedure is employed, to be exact the recovered $d^{(i)}$ at each pixel does not correspond to the value at this pixel and it takes an average value of the neigboring region defined by vibration width in the image. As a result, recovered $d^{(i)}$ has a correlation with the values in the neigboring region. The spatial extent of this correlation depends also on the depth value, and from the begining, $d^{(i)}$ has to be treated as the variable having such a correlation. We consider this as a future subject.

In this study, we assume that optical flow is very small, and hence, observation errors of f_t, f_x and f_y, which are calculated by finite difference, are small. Additionally, equation error is also small, and therefore we can assume that error having no relation with f_t, f_x and f_y is added to the whole gradient equation. From this consideration, we assume that $f_t^{(i,j)}$ is a Gaussian random variable with mean 0 and variance σ_o^2, and $f_x^{(i,j)}$ and $f_y^{(i,j)}$ have no error

$$\begin{aligned} p(f_t^{(i,j)} | d^{(i)}, r^{(j)}, \sigma_o^2) &= \frac{1}{\sqrt{2\pi}\sigma_o} \\ &\times \exp\left\{ -\frac{\left(f_t^{(i,j)} + f^{r(i,j)} + f^{u(i,j)} d^{(i)}\right)^2}{2\sigma_o^2} \right\}. \end{aligned} \qquad (6)$$

On the other hand, we also assume that $r^{(j)}$ is a 3-dimensional Gaussian random variable with mean 0 and variance-covariance matrix $\sigma_r^2 I$, where I indicates a 3×3 unit matrix

$$p(r^{(j)} | \sigma_r^2) = \frac{1}{(\sqrt{2\pi}\sigma_r)^3} \exp\left\{ -\frac{r^{(j)\top} r^{(j)}}{2\sigma_r^2} \right\}. \qquad (7)$$

From both models, the joint distribution of $\{f_t^{(i,j)}\}$ and $\{r^{(j)}\}$ is formulated as follows:

$$
\begin{aligned}
& p(\{f_t^{(i,j)}\},\{r^{(j)}\}|\Theta) \\
&= \prod_{i=1}^{N}\prod_{j=1}^{M} p(f_t^{(i,j)}|d^{(i)},r^{(j)},\sigma_o^2)\prod_{j=1}^{M} p(r^{(j)}|\sigma_r^2) \\
&= \frac{1}{(2\pi)^{M(N+3)/2}\sigma_o^{MN}\sigma_r^{3M}} \\
&\quad \times \exp\left\{ -\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(f_t^{(i,j)} + w^{(i,j)\top} r^{(j)}\right)^2}{2\sigma_o^2} \right. \\
&\qquad \left. -\frac{\sum_{j=1}^{M} r^{(j)\top} r^{(j)}}{2\sigma_r^2} \right\}, \qquad (8)
\end{aligned}
$$

$$
\begin{aligned}
w^{(i,j)} &= \begin{pmatrix} f_x^{(i,j)} x^{(i)} y^{(i)} + f_y^{(i,j)}(1 + y^{(i)^2}) \\ -f_x^{(i,j)}(1 + x^{(i)^2}) - f_y^{(i,j)} x^{(i)} y^{(i)} \\ f_x^{(i,j)} y^{(i)} - f_y^{(i,j)} x^{(i)} \end{pmatrix} \\
&\quad + Z_0 d^{(i)} \begin{pmatrix} f_y^{(i,j)} \\ -f_x^{(i,j)} \\ 0 \end{pmatrix} \\
&\equiv w_0^{(i,j)} + Z_0 d^{(i)} w_d^{(i,j)}, \qquad (9)
\end{aligned}
$$

where $\Theta = \{\{d^{(i)}\},\sigma_o^2,\sigma_r^2\}$. Additionally, the posterior distribution of $\{r^{(j)}\}$ is

$$
p(\{r^{(j)}\}|\{f_t^{(i,j)}\},\Theta) = \frac{p(\{r^{(j)}\},\{f_t^{(i,j)}\}|\Theta)}{p(\{f_t^{(i,j)}\}|\Theta)}, \quad (10)
$$

and this can be arranged as the following Gaussian distribution

$$
\begin{aligned}
p(\{r^{(j)}\}|\{f_t^{(i,j)}\},\Theta) &= \frac{1}{\sqrt{(2\pi)^{3M}\prod_{i=1}^{M}\det V_r^{(j)}}} \\
\times \exp&\left\{ -\frac{1}{2}\sum_{j=1}^{M}\left(r^{(j)} - r_m^{(j)}\right)^{\top} V_r^{(j)^{-1}} \left(r^{(j)} - r_m^{(j)}\right) \right\}, \qquad (11)
\end{aligned}
$$

where

$$
r_m^{(j)} = -\frac{1}{\sigma_o^2} V_r^{(j)} \sum_{i=1}^{N} f_t^{(i,j)} w^{(i,j)}, \qquad (12)
$$

$$
V_r^{(j)} = \left(\frac{1}{\sigma_o^2}\sum_{i=1}^{N} w^{(i,j)} w^{(i,j)\top} + \frac{1}{\sigma_r^2} I \right)^{-1}. \qquad (13)
$$

3.4 Computation Algorithm

In order to determine Θ as a maximum likelihood estimator and to determine $\{r^{(j)}\}$ as a MAP estimator, we apply the EM algorithm by treating $\{\{f_t^{(i,j)}\},\{r^{(j)}\}\}$ as a complete data and $\{r^{(j)}\}$ as a missing data.

The log likelihood function of the complete data $l_c(\Theta)$ is derived from Eq. 8 as

$$
\begin{aligned}
l_c(\Theta) &= \text{Const.} - \frac{MN}{2}\ln\sigma_o^2 - \frac{3M}{2}\ln\sigma_r^2 \\
&- \frac{1}{2\sigma_o^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(f_t^{(i,j)} + w^{(i,j)\top} r^{(j)}\right)^2 - \frac{1}{2\sigma_r^2}\sum_{j=1}^{M} r^{(j)\top} r^{(j)} \\
&= \text{Const.} - \frac{MN}{2}\ln\sigma_o^2 - \frac{3M}{2}\ln\sigma_r^2 \\
&\quad - \frac{1}{2\sigma_o^2}\sum_{j=1}^{M}\left\{ \sum_{i=1}^{N} f_t^{(i,j)^2} + 2\left(\sum_{i=1}^{N} f_t^{(i,j)} w^{(i,j)\top}\right) r^{(j)} \right. \\
&\quad \left. + \mathrm{tr}\left[\left(\sum_{i=1}^{N} w^{(i,j)} w^{(i,j)\top}\right) r^{(j)} r^{(j)\top}\right] \right\} \\
&\quad - \frac{1}{2\sigma_r^2}\sum_{j=1}^{M} \mathrm{tr}\left(r^{(j)} r^{(j)\top}\right). \qquad (14)
\end{aligned}
$$

In the EM algorithm, the E step and the M step are mutually repeated until they converge. At first, in the E step, the conditional expectation of the log likelihood with observing $\{f_t^{(i,j)}\}$, which is called Q function, is computed. In the Q function, the estimated value $\hat{\Theta}$ is used for the parameters values in the conditional distribution. In the following, the values computed using $\hat{\Theta}$ are indicated as $\hat{\cdot}$. Taking expectation of Eq. 14 results in expectation of the terms containing $\{r^{(j)}\}$, and using

$$
\mathrm{E}\left[r^{(j)}\right] \equiv r_m^{\widehat{(j)}} \qquad (15)
$$

and

$$
\mathrm{E}\left[r^{(j)} r^{(j)\top}\right] \equiv R^{\widehat{(j)}} = V_r^{\widehat{(j)}} + r_m^{\widehat{(j)}} r_m^{\widehat{(j)}\top}, \qquad (16)
$$

and ignoring constant value, the Q function becomes

$$
\begin{aligned}
Q(\Theta) &= -\frac{MN}{2}\ln\sigma_o^2 - \frac{3M}{2}\ln\sigma_r^2 \\
&- \frac{1}{2\sigma_o^2}\sum_{j=1}^{M}\left\{ \sum_{i=1}^{N} f_t^{(i,j)^2} + 2\left(\sum_{i=1}^{N} f_t^{(i,j)} w^{(i,j)\top}\right) r_m^{\widehat{(j)}} \right. \\
&\left. + \mathrm{tr}\left[\left(\sum_{i=1}^{N} w^{(i,j)} w^{(i,j)\top}\right) R^{\widehat{(j)}}\right] \right\} - \frac{1}{2\sigma_r^2}\sum_{j=1}^{M} \mathrm{tr} R^{\widehat{(j)}}. \\
& \qquad (17)
\end{aligned}
$$

In the M step, Θ is updated so as to maximize the Q function. We rewrite Eq. 17 as follows:

$$
\begin{aligned}
Q(\Theta) &= -\frac{MN}{2}\ln\sigma_o^2 - \frac{3M}{2}\ln\sigma_r^2 \\
&\quad - \frac{1}{2\sigma_o^2}\hat{F}(\{d^{(i)}\}) - \frac{1}{2\sigma_r^2}\hat{G}. \qquad (18)
\end{aligned}
$$

From this representation, σ_o^2 and σ_r^2 can be updated as

$$
\sigma_o^2 = \frac{\hat{F}(\{d^{(i)}\})}{MN}, \qquad \sigma_r^2 = \frac{\hat{G}}{3M}. \qquad (19)
$$

Additionally, $\{d^{(i)}\}$ can be also updated as follows:

$$d^{(i)} = -\frac{\sum_{j=1}^{M}\left\{ f_t^{(i,j)} {w_d^{(i,j)}}^{\top} \widehat{r_m^{(j)}} + \mathrm{tr}\left(B^{(i,j)}\widehat{R^{(j)}}\right)\right\}}{Z_0 \sum_{j=1}^{M} \mathrm{tr}\left(A^{(i,j)}\widehat{R^{(j)}}\right)}, \tag{20}$$

where the matrices $A^{(i,j)}$ and $B^{(i,j)}$ are defined as

$$A^{(i,j)} \equiv w_d^{(i,j)} {w_d^{(i,j)}}^{\top}, \tag{21}$$

$$B^{(i,j)} \equiv \frac{w_d^{(i,j)} {w_0^{(i,j)}}^{\top} + w_0^{(i,j)} {w_d^{(i,j)}}^{\top}}{2}. \tag{22}$$

4 NUMERICAL EVALUATIONS

To confirm the effectiveness of the proposed method, 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 128×128 pixels. In Fig. 3(b), the vertical axis indicates the depth Z and the horizontal axes means (x, y) in the image plane.

In our model, pairs of two successive images are assumed to be used in turn to calculate f_t. For this model, we have to adjust the correlation between successive rotations in order to keep the movement range at each image position in a certain local region, otherwise each position may move divergently as a random walk model. In these evaluations, to simplify the procedures, each rotation value was generated as a Gaussian independent random variable by computer, and the pairs to define f_t were taken as the original image and each successive image. Additionally, in order to firstly justify our algorithm for the assumed statistical models, we computed $\{f_t\}$ using Eq. 5 with the true value of r and $\{d\}$ and use them for depth recovery.

Figure 4 shows examples of the recovered depth map. The random value of each component of r was generated as a Gaussian random variable with mean 0 and deviation 0.01 [rad./frame]. Under this condition, the mean magnitude of optical flow took the value between one and two pixels. These results shown in Fig. 4 were calculated from $\{f_t\}$ having noise. A Gaussian random values with mean 0 and deviation corresponding to 1% of the deviation of the true $\{f_t\}$ were added to the true $\{f_t\}$. The initial value of both σ_o^2 and σ_r^2 was 1.0×10^{-2} as an arbitrary value, and $\{d\}$ was assumed initially as a plane of $Z = 9.0$. By varying the value of M corresponding to the number of sets $\{f_t\}$ between 100 and 800, we confirmed the effectiveness of collective utilization of many observations. The error maps of the recovered depth maps are also shown in Fig. 5. Additionally, the RMSEs of the recovered depth with respect to the noise deviation of $\{f_t\}$ are shown in Fig. 6. The outliers of the recovered depth taking the value below 6 or over 12 were excluded for evaluation of the RMSEs. From these results, we can conclude that the observations collection works well for accurate recovery.

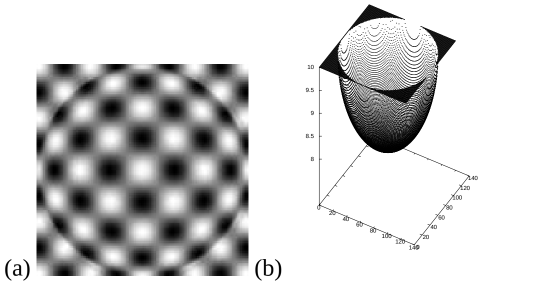


Figure 3: Example of the data used in the experiments: (a) artificial image used as an original image for making the successive images; (b) true depth map used for generating the images.

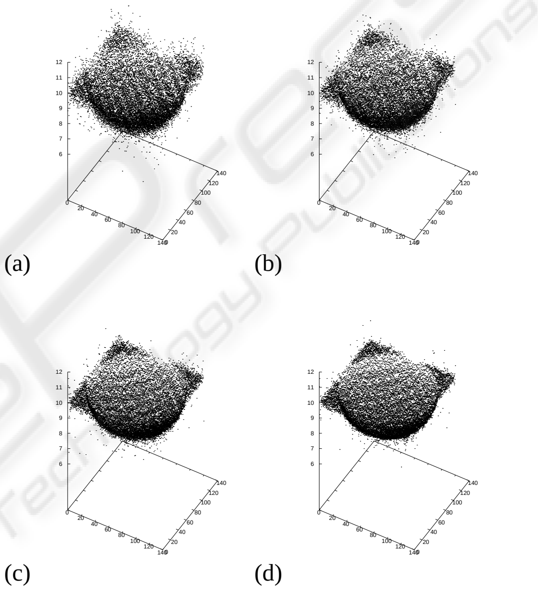
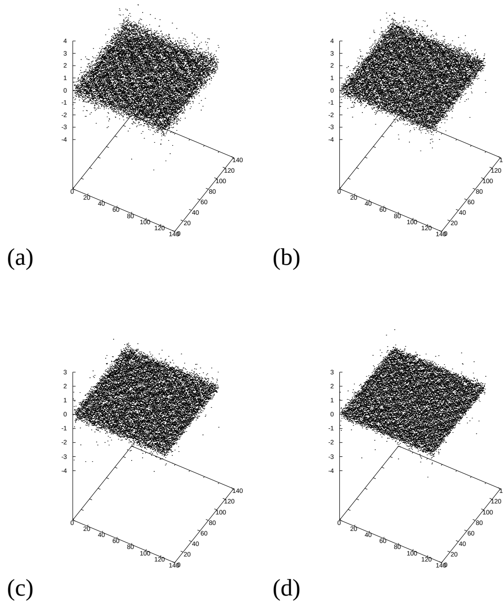


Figure 4: Stability of the proposed model for 1% noise of f_t: (a) $M = 100$; (b) $M = 200$; (c) $M = 400$; (d) $M = 800$.

5 CONCLUSIONS

In this paper, we propose a depth perception model with fixational eye movements. Especially for tremor, we construct a computation algorithm which recovers depth at each pixel collectively using multiple images over the period of one drift. Since this algorithm treats



(a) (b)

(c) (d)

Figure 5: Error map corresponding to the recovered depth shown in Fig. 4: (a) $M = 100$; (b) $M = 200$; (c) $M = 400$; (d) $M = 800$.

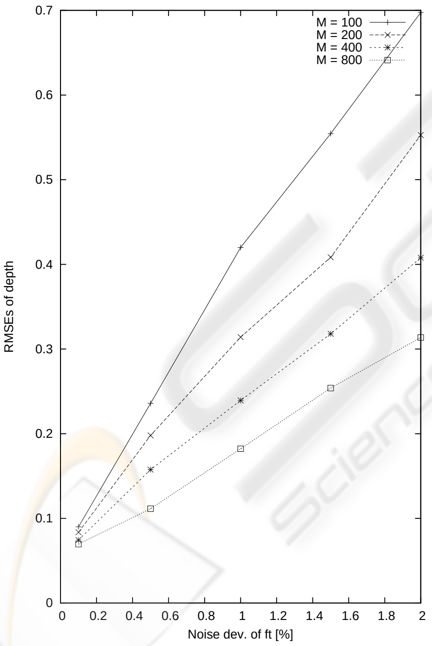


Figure 6: RMSEs of recovered depth with respect to noise deviation of f_t by varying M.

small changes of image brightness pattern, the linear approximation error contained in the gradient equation becomes small. Moreover, because one depth map corresponding to multiple successive images is recoved, the bad influence of observation errors can be reduced.

In future, in order to get an accurate depth map with small successive images, we are going to examine a model in which depth values in the local region are assumed to be constant or to have spatial correlation. Additionally, we have to construct whole algorithm based on fixational eye movement and binocular stereopsis, and have to show the effectiveness of the algorithm through the real image experiments.

REFERENCES

Ando, S., Ono, N., and Kimachi, A. (2002). Involuntary eye-movement vision based on three-phase correlation image sensor. In *proc. 19th Sensor Symposium*, pages 83–86.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data. *J. Roy. Statist. Soc. B*, 39:1–38.

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

Lazaros, N., Sirakoulis, G. C., and Gasteratos, A. (2008). Review of stereo vision algorithm: from software to hardware. *Int. J. Optomechatronics*, 5(4):435–462.

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

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

Tagawa, N., Kawaguchi, J., Naganuma, S., and Okubo, K. (2008). Direct 3-d shape recovery from image sequence based on multi-scale bayesian network. In *proc. ICPR '08*, pages CD–ROM.