Chapter

Stereoscopic Calculation Model Based on Fixational Eye Movements

Norio Tagawa

Abstract

Fixational eye movement is an essential function for watching things using the retina, which has the property of responding only to changes in incident light. However, since the rotation of the eyeball causes the translational movement of the crystalline lens, it is possible in principle to recover the depth of the object from the moving image obtained in this way. We have proposed two types of depth restoration methods based on fixation tremor; differential-type method and integral-type method. The first is based on the change in image brightness between frames, and the latter is based on image blurring due to movement. In this chapter, we introduce them and explain the simulations and experiments performed to verify their operation.

Keywords: motion stereoscopic, fixational eye movements, differential-type method, integral-type method, optical flow, gradient equation, image blur

1. Introduction

When humans stare at a target, an irregular involuntary movement called fixational eye movements occur [1]. The human retina can maintain reception sensitivity by finely vibrating the image of the target on the retina, so in order to see something, first fixation motion is required. It has been reported that the vibrations may work not only as such the intrinsic function to preserve photosensitivity but also as an assistance in image analysis, the mechanism of which can be interpreted as an instance of stochastic resonance (SR) [1]. SR is inspired by biology, more specifically by neuron dynamics [2], and based on it, the Dynamic Retina (DR) [3] and the Resonant Retina (RR) [4], which are new vision devices taking advantage of random camera vibrations, were proposed for contrast enhancement and edge detection respectively. It has been reported that the movement of the retinal image due to fixation eye movements can be an unconscious clue to depth perception, and an actual vision system based on fixational eye movements has been proposed [5].

On the other hand, binocular stereopsis is vigorous and plays an essential role in depth perception of a human vision system [6]. In general, binocular stereopsis detects relatively large disparities, hence it can recognize high accurate depth. However this causes an occlusion problem, and a lot of solutions of it have been proposed. Wang et al. have proposed a local detector for occlusion based on deep learning [7]. In [8], a robust depth restoration method has been proposed that integrates line-field imaging technology that simultaneously observes multiple angle views with stereo vision. Therefore, we expect that primitive depth information detected by fixational eye movements can be used to solve occlusion for binocular stereopsis. There is a concern that the accuracy of depth restoration by a small camera motion is lower than that of stereo vision. Even so, it is expected that erroneous correspondence due to the existence of occlusion can be reduced by using the depth information from fixational eye movements for the correspondence problem in stereo vision.

In monocular stereoscopic vision, "structure from motion (SFM)" has been the most widely studied, and many remarkable results have been reported. SFM has various calculation principles. To achieve spatially dense depth recovery with high computational efficiency, a method based on the gradient equation that expresses the constraint between the spatiotemporal derivative values of image intensity and the movement on the image is effective [9–11]. It should be noted that for such a gradient method, there is an appropriate size of movement to recover the correct depth. Since the gradient equation holds perfectly for small motions, the error in the equation cannot be ignored for very large motions. On the contrary, in the case of small movement, the motion information is buried in the observation error of the spatiotemporal derivative in intensity.

Adaptation of the frame rate is required to make the motion size suitable for the gradient method. We have proposed a method that does not require a variable frame rate based on multi-resolution decomposition of images, but it requires high computational cost [12]. Therefore, we focus on small movements with an emphasis on avoiding equation errors in the gradient method. Then, in order to solve the above signal-to-noise ratio (S/N) problem that occurs with small movements, many observations are collected and used all at once [13, 14]. In such a strategy, it is desirable that the direction and size of the motion take different values. From the above discussion, we examined a depth perception model based on fixational eye movement and gradient method. Fixational eye movements are divided into three types: microsaccades, drifts, and tremors. As the first report of our attempt, we focused on tremor, the smallest of the three types. In the next step, we plan to use drift and microsaccade analogies for further progress. Using a lot of images captured with random small motions of camera, which consists of three-dimensional (3-D) rotations imitating fixational eyeball motions [1], 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. Since the difference between the center of the three-dimensional rotation and the lens center generates a translational motion of the lens center, depth information can be obtained from these images. Simulations with artificial images confirm that the proposed method works effectively when the observed noise is an actual sample of a theoretically defined noise model.

However, if the wavelength of the main luminance pattern is small compared to the size of the motion in the image, aliasing will occur and the gradient equation will be useless. In other words, the methods of [13, 14] cannot be applied. To avoid the above problem, we proposed a new scheme based on the integral form that also used the analogy of fixational eye movement [15, 16]. Add up the many images generated by the above method to get one blurry image. The degree of blur is a function of pixel position and also depends on the depth value of each pixel. That is, the difference in the degree of image blur indicates the depth information. Based on the proposed scheme, the spatial distribution of the image blur is effectively estimated using the blurred image and the original image without blur. By modeling the small 3-D rotation of the camera as a Gaussian random variable, the depth map can be calculated analytically from this blur distribution.

Several depth recovery methods using motion-blur have been already proposed, but those use the blur caused by definite and simple camera motions. For example,

smooth depth tends to be recovered from the recognized smooth motion blur from **Figure 8(c)**, it can be confirmed that the smoothness constraint of Eq. (23) is an obstacle to the reduction of RMSE.

5. Real image experiments of differential-type method

5.1 Selective use of image pairs to improve accuracy

When applying the difference-type method to an actual image and checking the actual performance, the performance was improved by selecting the image pair used for depth restoration. We have adopted a scheme that excludes image pairs that are expected to have large approximation errors in the gradient equation on a pixel-by-pixel basis. We can use the inner product of the spatial gradient vectors of consecutive image pairs to select image pairs that do not cause aliasing problems. For each pixel, the image pairs of which the sign of the inner product $\boldsymbol{f}_s^{(i,j)^{\mathrm{T}}}\boldsymbol{f}_s^{(i,j-1)}$ is negative are discarded. It is noted that $\boldsymbol{f}_s^{(i,j)} = \left[f_x^{(i,j)}, f_y^{(i,j)}\right]^{\mathrm{T}}$.

In the next step, from the image pairs remained by the above decision, we additively select the suitable image pairs at each pixel by estimating the amount of the higher order terms included in the observation of f_t. f_t is exactly represented as follows:

$$f_t = -f_x v_x - f_y v_y - \frac{1}{2}\left\{ f_{xx} v_x^2 + f_{yy} v_y^2 + 2 f_{xy} v_x v_y \right\} + \cdots. \tag{28}$$

After discarding a bad image pair, the higher-order terms can be considered small. In this case, the quadratic term in Eq. (28) can be estimated for each pixel i as follows:

$$-\frac{1}{2}\left\{\left(f_x^{(i,j)} - f_x^{(i,j-1)}\right)v_x^{(i,j)} + \left(f_y^{(i,j)} - f_y^{(i,j-1)}\right)v_y^{(i,j)}\right\}. \tag{29}$$

We can define a measure for estimating the equation error as the ratio of this higher order term to the first order term.

$$J = \frac{\left|\left(f_x^{(i,j)} - f_x^{(i,j-1)}\right)v_x^{(i,j)} + \left(f_y^{(i,j)} - f_y^{(i,j-1)}\right)v_y^{(i,j)}\right|}{2\left|f_x^{(i,j)} v_x^{(i,j)} + f_y^{(i,j)} v_y^{(i,j)}\right|}. \tag{30}$$

This measurement depends on the direction of the optical flow but is invariant with respect to the amplitude of the optical flow. To calculate the value of J, we need to know the true value of the optical flow. By examining the details of J, even if the difference of the spatial gradient $\boldsymbol{f}_s^{(i,j)} - \boldsymbol{f}_s^{(i,j-1)}$ is large, when the direction of $\boldsymbol{f}_s^{(i,j)} - \boldsymbol{f}_s^{(i,j-1)}$ is perpendicular to that of optical flow, the equation error becomes small. Therefore, the value $\left|\boldsymbol{f}_s^{(i,j)} - \boldsymbol{f}_s^{(i,j-1)}\right| / \left|\boldsymbol{f}_s^{(i,j)}\right|$ can be used as the worst value. In the following, the image pairs for which $\left|\boldsymbol{f}_s^{(i,j)} - \boldsymbol{f}_s^{(i,j-1)}\right| / \left|\boldsymbol{f}_s^{(i,j)}\right|$ is less than the certain threshold value are selected at each pixel to be used for depth recovery.

5.2 Camera system implementation

We built the camera hardware system for examining the practical performance of our camera model shown in **Figure 1**. The implemented camera system is shown in **Figure 9**.

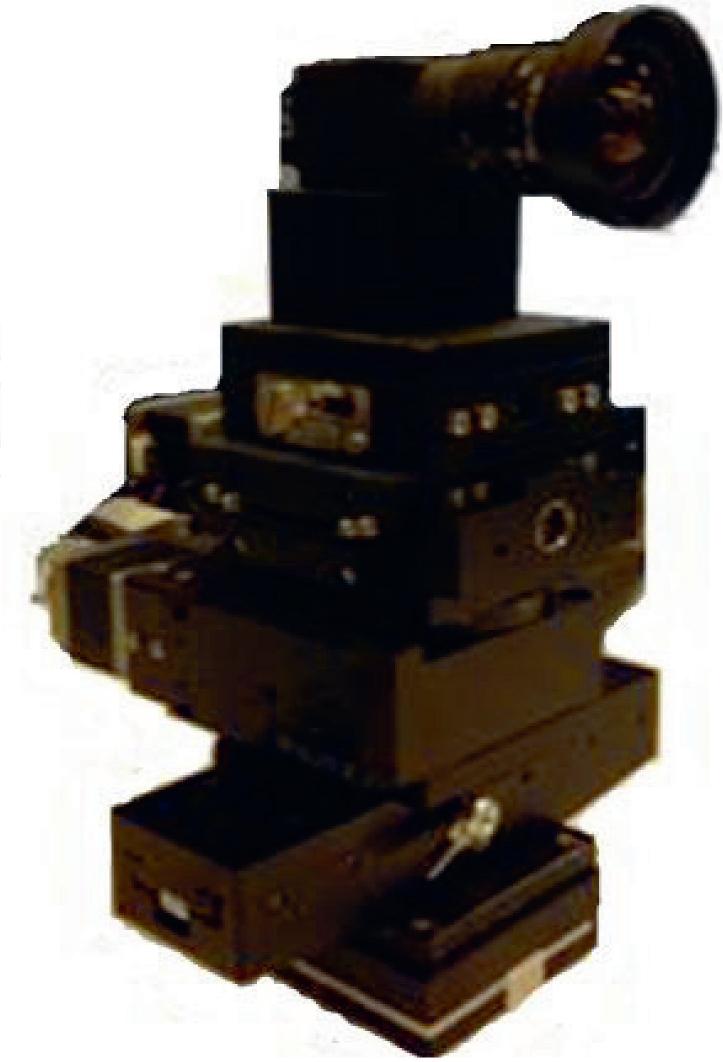

Figure 9.
Camera system implemented for tremor rotations.

The camera system can be rotated around the horizontal axis i.e. X axis and around the vertical axis, i.e. Y axis. The rotation around the optical direction, i.e. Z direction, cannot be performed, which is not needed to gain the depth information. The parameters of the system are shown as follows: focal length is $2.8 - 5.0$ mm, image size is $1{,}200 \times 1{,}600$ pix., movable widths are 360 deg. for X axis and $(-10, +10)$ deg. for Y axis, and drivable minimum units are 1 pulse = 0.01 deg. for X-axis and 1 pulse = 0.00067 deg. for Y-axis.

5.3 Experimental results

In this section, we explain the results of the experiments using the real images captured by the developed camera system [22]. Our camera system has a parallel stereo function. That is, the camera can be moved laterally by the slide system. Prior to the experiment, we calibrated the camera's internal parameters, including focal length and Z_0, using the method in [23] and stereo calculations. The image used in the experiment is grayscale, consists of 256×256 pixels, and is 8-bit digitized. An example is shown in **Figure 10(a)**. The true inverse depth of the target object is shown in **Figure 10(b)**. It was measured in parallel stereo above using a two-plane model. In this figure, the horizontal axis shows the position in the image plane, and the vertical axis shows the inverse depth in units of focal length. We captured 100 images. The maximum number of iterations of the MAP-EM

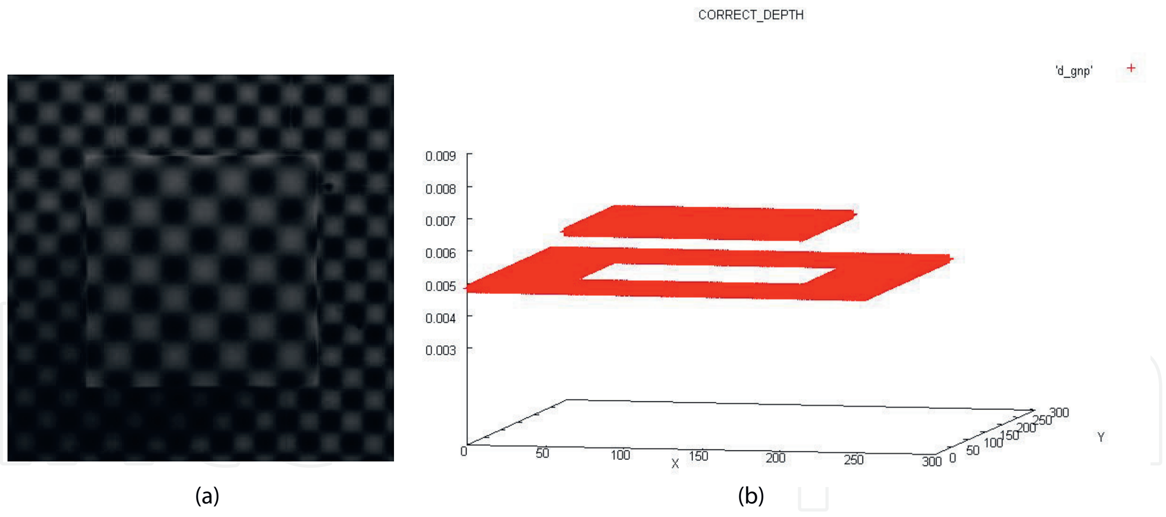

Figure 10.
Data for experiments: (a) example of captured image, (b) true inverse depth of object (reprinted from [22]).

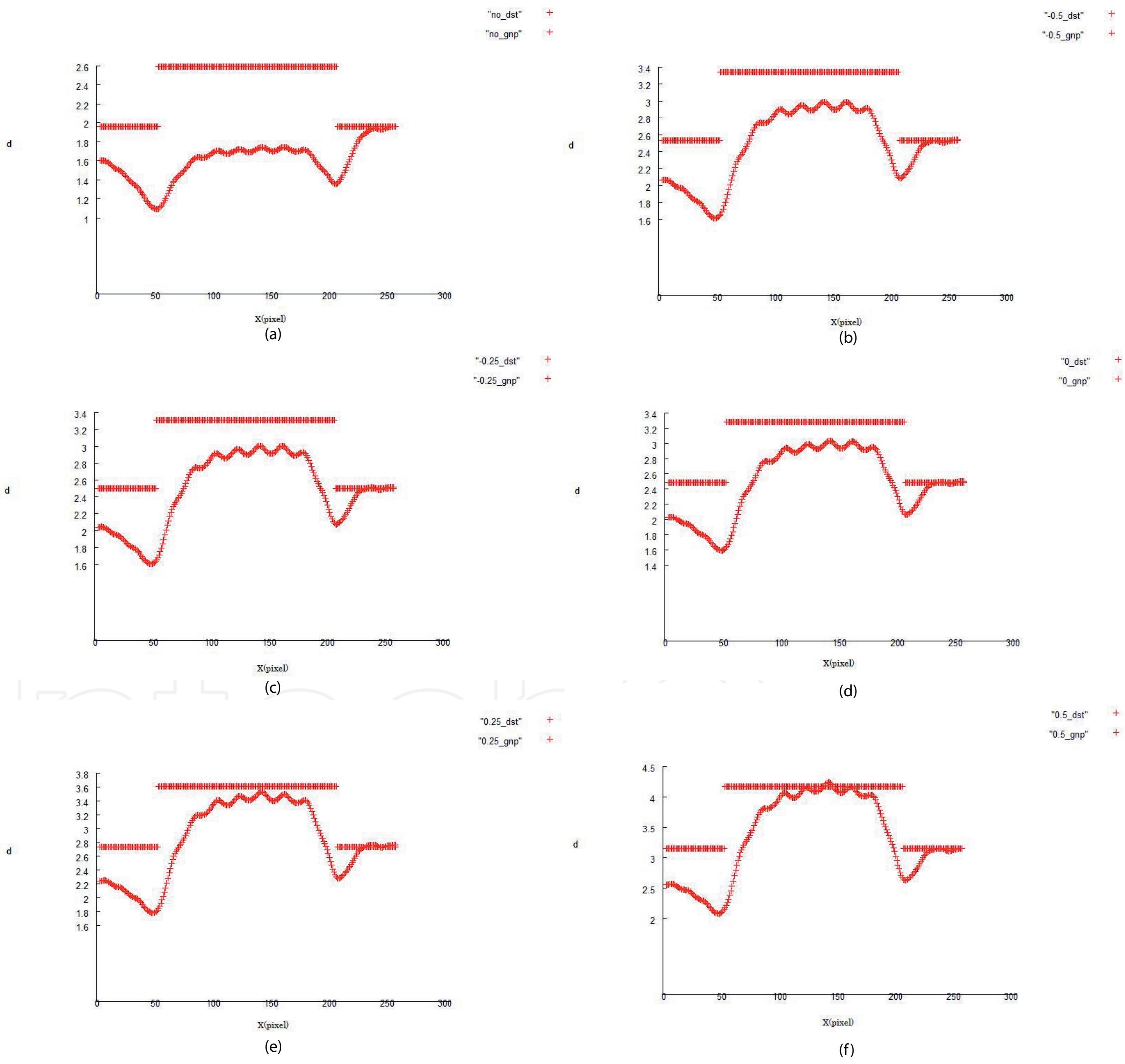

Figure 11.
Profiles of cross-section of recovered inverse depth: (a) all image pairs are used (100%), (b) threshold ×1.5(94% image pairs were used), (c) threshold ×1.25(86%), (d) threshold ×1(68%), (e) threshold ×0.75(62%), (f) threshold ×0.5(62%) (reprinted from [22]).

algorithm was set to 600. Within this number of iterations, the iterations of almost all experiments converged. σ_d^2 is heuristically determined. The average value of $|f_s^{(i,j)} - f_s^{(i,j-1)}|/|f_s^{(i,j)}|$ explained in the previous section with respect to all pixels was defined for each image pair as a standard magnification (×1) of the threshold for selecting the suitable image pairs. Namely, by decreasing the threshold magnification, we can discard more image pairs. Conversely, by increasing the magnification, many image pairs can be used for recovery. Because of the limit of pages, we only show the results with $\sigma_r^2 = 2.64 \times 10^{-2}$ by which the average of the optical flow's amplitude approximately coincides with $\lambda/4$.

Figure 11 shows the result of the recovered depth for each threshold set as a constant multiple of the reference value. We also looked at the results using all image pairs. From these results, it can be confirmed that by reducing the magnification, inappropriate image pairs can be discarded and the accuracy of depth recovery is improved. The percentage shown in the caption of the figure shows the number of image pairs used for recovery, which is determined in conjunction with the change in threshold.

6. Conclusions

In this chapter, we introduced a depth recovery algorithms that uses large number of images with small movements by using camera motion that simulates fixational eye movements, especially the tremor component. The algorithms can be divided into a differential-type and an integral-type. For the differential-type, it is desirable that the movement on the image is relatively small with respect to the texture pattern of the surface to be imaged, and conversely, for the integral-type, it is appropriate to apply it to a fine texture compared to the movement on the image. Therefore, ideally, the development of a depth recovery system in which both schemes function adaptively and selectively according to the target texture is the most important task in the future.

A detailed technical issue is to automatically determine the parameters that control the smoothness of the depth. This can be achieved by considering all unknowns as stochastic variables and formulating them in the variational Bayesian framework. As for the integration method, since the resolution of the recovered depth is low in principle, it is possible to consider a composite type in which the differential-type is applied again and refinement is performed on the result obtained by the integral-type.

So far, we have considered a method that assumes only tremor, but in the future, we are planning to study camera motion that also simulates drift and microsaccade. In the method for drift component, it is necessary to extend the method based on tremor to the online version, and then update the depth estimate while advancing the tracking of the target as time series processing. When using microsaccades, it is necessary to handle large movements between frames. Therefore, based on the correspondence of feature points, sparse but highly accurate depth restoration can be expected. Drift itself does not have much merit in its use, but it plays an important role in generating microsaccades. As described above, we believe that an interesting system can be realized by comprehensively using the three components.

On the other hand, stereoscopic vision and motion stereoscopic vision are difficult to handle objects with few textures. In [24], we proposed a stereo system that considers shading information. The projected images to both cameras are calculated by computer graphics technique while changing the depth estimation value. The depth is determined so that the generated image matches the image observed by each camera. As a result, the association between images is indirectly realized. By introducing this method, it becomes possible to handle textureless objects. We aim to develop a comprehensive depth restoration method, including the multi-resolution processing proposed in [12]. In another scheme that deals with the textureless region in stereo vision, the region where the depth value is constant or changes smoothly, called the support region, is adaptively determined [25]. We will also consider whether the relationship between image changes due to tremor and microsaccade can be used for adaptive determination of this support region.

In recent years, many realizations of stereoscopic vision and motion stereoscopic vision by deep learning have been reported [26–28]. And the relationship with the conventional method based on mathematical formulas is often questioned. The deep learning method is hampered by the addition of a large number of images and annotations to them. Although unsupervised learning is often devised, the solution is often limited. Therefore, even if the conventional method is rather complicated and takes time, if a method capable of more precise depth recovery is constructed, it can be used for annotation calculation of deep learning. This can be understood as copying the conventional method to deep neural network (DNN). DNN takes time to learn, but has the advantage of being able to infer at high speed. In this way, it is important that both schemes develop in a two-sided relationship.

Appendix

Here, the method of calibrating the axis of rotation is explained using **Figure 12**. Let a point in 3-D space be $\boldsymbol{X}_1 = [X_1, Y_1, Z_1]^{\mathrm{T}}$ in the coordinate system before camera rotation and $\boldsymbol{X}_2 = X_2, Y_2, Z_2]^{\mathrm{T}}$ in the coordinate system after rotation, and the coordinates of the corresponding points on the image be $\boldsymbol{x}_1 = \left[x_1, y_1, z_1\right]^{\mathrm{T}}$ and $\boldsymbol{x}_2 = \left[x_2, y_2, z_2\right]^{\mathrm{T}}$, respectively. Similarly, the optical axes before and after rotation are $\boldsymbol{z}_1^1 = [0, 0, 1]^{\mathrm{T}}$ and $\boldsymbol{z}_2^2 = [0, 0, 1]^{\mathrm{T}}$, respectively. If the rotation is taken around the X-axis, the rotation matrix is given by the following equation.

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}. \tag{31}$$

The translation $\boldsymbol{T}$ of the lens center generated by this rotation is given by the following equation in the coordinate system before rotation.

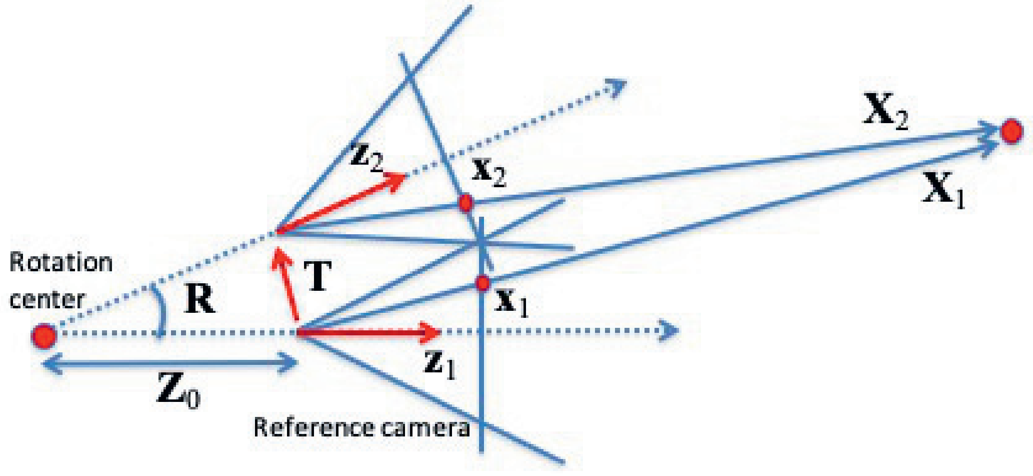

Figure 12.
Explanation of rotation axis calibration.

$$\boldsymbol{T}^1 = Z_0\, z_2^1 - Z_0\, z_1^1 = Z_0(\boldsymbol{R} - \boldsymbol{I})\, z_1^1 \equiv Z_0\, \boldsymbol{S} z_1^1, \tag{32}$$

where z_2^1 represents the optical axis after rotation in the coordinate system before rotation. In addition, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the following relationship.

$$\boldsymbol{X}_2 = \boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{X}_1 - \boldsymbol{T}^1\right) \rightarrow \boldsymbol{R}\boldsymbol{X}_2 = \boldsymbol{X}_1 - \boldsymbol{T}^1 \tag{33}$$

Furthermore, by substituting the relation of $\boldsymbol{x}_1 = \boldsymbol{X}_1^1/Z_1$, $\boldsymbol{x}_2 = \boldsymbol{X}_2^2/Z_2$ into Eq. (33), the following equation is obtained.

$$Z_2 \boldsymbol{R}\boldsymbol{x}_2 = \boldsymbol{X}_1 - Z_0\, \boldsymbol{S} z_1^1. \tag{34}$$

By expressing this equation in terms of components and organizing it, the following two equations are derived.

$$Z_2\left(y_2 \cos\theta - \sin\theta\right) = Y_1 + Z_0 \sin\theta, \tag{35}$$

$$Z_2\left(y_2 \sin\theta + \cos\theta\right) = Z_1 - Z_0(\cos\theta - 1). \tag{36}$$

By substituting Eq. (35) into Eq. (36), the solution of Z_0 can be derived as follows:

$$Z_0 = \frac{Z_1\left(y_2 \cos\theta - \sin\theta\right) - Y_1\left(y_2 \sin\theta + \cos\theta\right)}{\sin\theta\left(y_2 \sin\theta + \cos\theta\right) + (\cos\theta - 1)\left(y_2 \cos\theta - \sin\theta\right)}. \tag{37}$$

Author details

Norio Tagawa
Graduate School of Systems Design, Tokyo Metropolitan University, Tokyo, Japan

*Address all correspondence to: tagawa@tmu.ac.jp

References

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

[2] Stemmler M. A single spike suffices: the simplest form of stochastic resonance in model neuron. Network: Computations in Neural Systems. 1996; 61:687–716.

[3] Propokopowicz P and Cooper P. The dynamic retina. Int. J. Computer Vision. 1995;16:191–204.

[4] Hongler M-O, de Meneses Y L, Beyeler A, and Jacot J. The resonant retina: Exploiting vibration noise to optimally detect edges in an image. IEEE Trans. Pattern Anal. Machine Intell. 2003;25:1051–1062.

[5] Ando S, Ono N, and Kimachi A. Involuntary eye-movement vision based on three-phase correlation image sensor. In: Proceedings on19th Sensor Symposium; 2002. p. 83–86.

[6] Lazaros N, Sirakoulis G-C, and Gasteratos A. Review of stereo vision algorithm: from software to hardware. Int. J. Optomechatronics. 2008;5:435–462.

[7] Wang J and Zickler T. Local detection of stereo occlusion boundaries. In: Proceedings on CVPR; 2019. p. 3818–3827.

[8] Liu F, Zhou S, Wang Y, Hou G, Sun Z, and Tan T. Binocular light-field: imaging theory and occlusion-robust depth perception application. IEEE Trans. Image Process. 2019;29: 1628–1640.

[9] Horn B K P and Schunk B. Determining optical flow. Artif. Intell. 1981;17:185–203.

[10] Simoncelli E P. Bayesian multi-scale differential optical flow. In: Jahne B, Haubecker H, Geibier P, editors. Handbook of Computer Vision and Applications.Academic Press; 1999. vol. 2. p. 397–422.

[11] Bruhn A and Weickert J. Locas/kanade meets horn/schunk: combining local and global optic flow methods. Int. J. Computer Vision. 2005;61:211–231.

[12] Tagawa N, Kawaguchi J, Naganuma S, and Okubo K. Direct 3-d shape recovery from image sequence based on multi-scale bayesian network. In: Proceedings on ICPR; 2008. p. CD–ROM.

[13] Tagawa N. Depth Perception model based on fixational eye movements using Bayesian statistical inference. In: Proceedings on ICPR; 2010. p. 1662–1665.

[14] Tagawa N and Alexandrova T. Computational model of depth perception based on fixational eye movements. In: Proceedings on VISAPP; 2010, p. 328–333.

[15] Tagawa N, Iida Y, and Okubo K. Depth perception model exploiting blurring caused by random small camera motions. In: Proceedings on VISAPP; 2012, p. 329–334.

[16] Tagawa N, Koizumi S, and Okubo K. Direct depth recovery from motion blur caused by random camera rotations imitating fixational eye movements. In: Proceedings on VISAPP; 2013, p. 177–186.

[17] Sorel M and Flusser J. Space-variant restoration of images degraded by camera motion blur. IEEE Trans. Image Processing. 2008;17:105–116.

[18] Paramanand C and Rajagopalan A N. Depth from motion and optical blur with unscented Kalman filter. IEEE Trans. Image Processing. 2012;21:2798–2811.

[19] Dempster A-P, Laird N-M, and Rubin D-B. Maximum likelihood from incomplete data. J. Roy. Statist. Soc. B. 1977;39:1–38.

[20] Green P.-J. On use of the Em algorithm for penalized likelihood estimation. J. Roy. Statist. Soc. B. 1990; 52:443–452.

[21] Poggio T, Torre V, and Koch C. Computational vision and regularization theory. Nature. 1985;317:314–319.

[22] Tsukada S, Ho Y, Tagawa N, and Okubo K. Accuracy improvement for depth from small irregular camera motions and its performance evaluation. In: Proceedings on ICIAR; 2015, p. 306–315.

[23] Zhang Z. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Machine Intell. 2000;22: 1330–1334.

[24] Wakabayashi K, Tagaw N, and Okubo K. Shape from multi-view images based on image generation consistency. In: Proceedings on VISAPP; 2013, p. 334–340.

[25] Wu W, Zhu H, Yu S, and Shi J. Stereo matching with fusing adaptive support weights. IEEE Access. 2019;7: 61960–61974.

[26] Laina I, Rupprecht C, Belagiannis V, Tombari F, and Navab N. Deeper depth prediction with fully convolutional residual networks. In: Proceedings on 3D Vision;2016, p. 239–248.

[27] Ummenhofer B, Zhau H, Uhrig J, Mayer N, Ilg E, Dosovitskiy A, and Brox T. DeMoN: Depth and motion network for learning monocular stereo. In: Proceedings on CVPR;2017, p. 5038–5047.

[28] Chen P-Y, Liu A H, Liu Y-C, Wang Y-C F. Towards scene understanding: unsupervised monocular depth estimation with semantic-aware representation. In: Proceedings on CVPR;2019, p. 2624–2632.