Depth Estimation from Tilted Optics Blur by Using Neural Network

Hiroshi Ikeoka
Fukuyama University
Department of Computer Science, Faculty of Engineering
Hiroshima, Japan
ikeoka@fukuyama-u.ac.jp

Takayuki Hamamoto
Tokyo University of Science
Department of Electrical Engineering, Faculty of Engineering
Tokyo, Japan
hamamoto@ee.kagu.tus.ac.jp

***Abstract*—We have been investigating a depth estimation system for real-time usage such as automotive tasks. Conventional method with stereo camera is too sensitive to slight variations of baseline length. Additionally, it has occlusion problem. Conversely, the method that uses a monocular camera by focusing cannot provide a balance between wide-area estimation and real-time estimation. Therefore, we proposed a novel method that adopts tilted lens optics. Herein, our method can obtain depth values at each pixel from the sharpness ratio of only two tilted optics images; our system is consisted of monocular camera system with spectroscopic mirror. Our method uses the optic lens which has some wider angle of view. For that reason, it causes some estimation error based on the difference between the actual camera system and the optical theory. Herein, to reduce the error, we adopted the neural network to calculate the depth value from the blur values and the y-coordinate. In this paper, we report our depth estimation method from tilted optics blur by using neural network.**

Keywords—depth estimation; distance estimation; tilted optics; blur; defocus; neural network; deep learning

I. INTRODUCTION

In the recent automotive industry, it has come to be realized that the importance of active safety technology for preventing traffic accidents in addition to passive safety technology for damage reduction following the occurrence of traffic accidents. Depth estimation techniques are central to both technologies. In fact, collision avoidance devices that utilize various sensors have already started being used.

However, existing technology does not satisfactorily answer the requirements of automotive applications. For example, a millimeter wave method is commonly used, and it has sufficiently high accuracy for depth estimation [1]. However, this method obtains only one depth value corresponding to one sensing. Therefore, to cover a wide-area, there is a need to achieve wider angle using swing mechanism or phase difference of multi-sensors. Therefore, the millimeter wave method causes sometimes false detections such as automobiles in other lanes on the narrow road or at the corner road. Moreover, it is not easy to detect soft bodies such as human. As another example, the method with the ultrasonic wave which attenuate immediately in the air can only obtain the depth in near area [2]. Similarly, infrared method which is easy affected by the sunlight used in near area only; Also, the method with the laser-radar is extremely more expensive than the main body of vehicle. Accordingly, it will take a long time before it becomes widespread. As stated above, it is difficult to prevent collision by only the existing active sensing methods. Hence, it is needed to use a combination of vision systems; consequently, the cost of a device increases, and installation becomes more complicated.

Passive methods are exemplified by the stereo camera method, which uses visible light to obtain wide-area depth information [3,4,5]. When using the stereo camera, it is important to maintain the base length between optic axes. Especially, unlike indoor robots of factory automation systems, the relationship between two axes in automotive applications is easy to be broken by disturbances such as vibration from tires, thermal expansion from temperature change. Hence, the stereo method requires a tough base mount and its drawback is that its device size is larger in spite of small size camera. Moreover, the stereo method requires two all in-focus images obtained using a camera with small lens aperture. This implies that the amount of incident light is small; this limitation reduces the merits of the method when used in dark areas and highways.

Conversely, monocular camera methods, such as the Depth-from-focus and the Depth-from-defocus methods [6,7,8]. Therefore, these methods are not suitable for deep depth estimation, because the disadvantages of these are unsatisfactory amount of incident light, device costs and durability. Additionally, the method with getting the targets depth by the pattern matching can detect the registered pattern only such as the back face of the cars.

In this paper, we propose a depth estimation method, which utilizes a monocular aperture camera with two tilted optics, two image sensors, and a spectroscopic mirror [9]. These optics make the plane of sharp focus (POF) lain and the depth of field (DOF) enlarged toward the depth direction (Fig 1). By using this method, it is possible to achieve depth estimation that satisfies needs such as 1) real-time processing, 2) wide estimation range, 3) satisfactory image brightness, and 4) robustness for use in automotive environments.

However, it causes some estimation error based on the difference between the actual camera system and the optical theory. Herein, to reduce the error, we propose to adopt the neural network to calculate the depth value from the sharpness values and the y-coordinate.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

II. TITLTED OPTICS IMAGING

In a situation with a tilted optics, the plane of the image sensor, the principal plane of the lens, the POF and DOF limits are not parallel. In this section, we describe the tilted optics which our proposed method adopts.

A. POF with a tilted optics

Let us now consider the situation schematically depicted in Fig.1: an image sensor that is placed on the left side of a horizontal axis and an optical unit (lens) that has a tilted angle, θ, with respect to the image sensor. Similarly, a POF is tilted, which is different from normal optics. The point (L_y, L_z) is the center of rotation for the lens, H is the distance between the first principal point and the center of rotation, and H' is the distance between the second principal point and the center of rotation. There are two light lines that reach the POF via each focal point, and one light line that reaches the POF via the two principal points. Hence, we can obtain a point of intersection of these light lines on the POF; the z and y coordinates are expressed with Y on the image sensor. Furthermore, by eliminating the Y from each coordinate, we can obtain the following POF expression:

$$y = \frac{1}{\sin\theta}\left(\cos\theta - \frac{f}{L_z}\right)z - \frac{L_z\cos\theta - L_y\sin\theta}{\sin\theta}. \quad (1)$$

At present, in order to simplify, we assumed that:

$$H = H' = 0. \quad (2)$$

When (2) is satisfied, an image sensor plane, a lens principal plane, and a POF all intersect at the following point on the y-axis:

$$\left(-\frac{L_z\cos\theta - L_y\sin\theta}{\sin\theta}, 0\right). \quad (3)$$

B. DOF with a tilted optics

By using tilted optics, both of the DOF limits shown by the dashed lines in Fig.1 are as well. If φ is the angle of the POF with respect to the ground, the gradient of the POF is:

$$\tan\varphi = \frac{1}{\sin\theta}\left(\cos\theta - \frac{f}{L_z}\right). \quad (4)$$

When we approximate the depth of focus with a constant value ϵ, the gradients of both DOF limits are derived as follows:

$$\tan\varphi_- = \tan\varphi - \frac{\epsilon}{f}\left(\frac{1}{\tan\theta} + \tan\varphi\right),$$
$$\tan\varphi_+ = \tan\varphi + \frac{\epsilon}{f}\left(\frac{1}{\tan\theta} + \tan\varphi\right). \quad (5)$$

Given that the DOF limits have gradients as in (5), the DOF limits are approximated by linear expression such as

$$y = \tan\varphi_- \cdot z - \frac{f}{\tan\theta} - L_z \cdot \tan\varphi_- + L_y,$$
$$y = \tan\varphi_+ \cdot z - \frac{f}{\tan\theta} - L_z \cdot \tan\varphi_+ + L_y. \quad (6)$$

III. DEPTH ESTIMATION

A. Sharpness curve with non-vertical POF and DOF

In this section, we describe the sharpness curve with non-vertical POFs. A plane that was characterized by white noise pattern with constant contrast was placed parallel to the image sensor, and we captured an image of the target plane by using non-vertical POF. We then calculated the sharpness value,

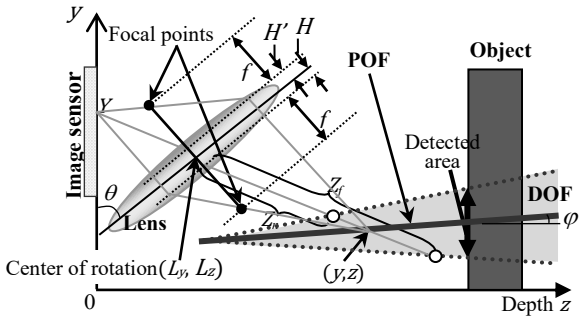
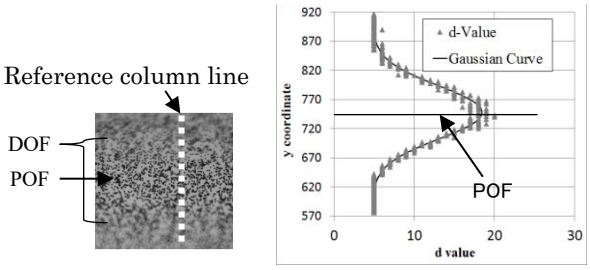


Fig.1 A titled POF and DOF when using a tilted lens.



(a) White noise image (b) Variance of *d*-value with LPF

Fig.2 White noise image when using tilted POFs.

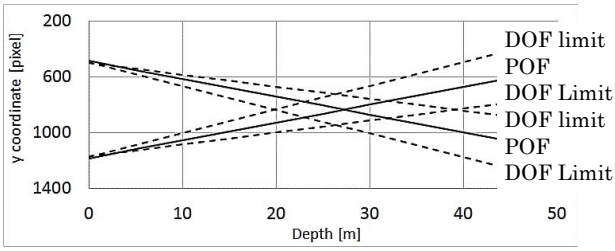


Fig.3 Arrangement of POFs and DOF limits.

which is the following *d*-value for this captured image, shown in Fig.2(a). The *d*-value is based on the derivation of a Laplacian:

$$d = \left|l_{i(x+1,y)} + l_{i(x-1,y)} - 2l_{i(x,y)}\right| + \left|l_{i(x,y+1)} + l_{i(x,y-1)} - 2l_{i(x,y)}\right|. \quad (7)$$

Here, l is the pixel's intensity and (x, y) are the pixel coordinates. When the *d*-value is high, the focal position is judged to be the in-focus position. Fig.2(b) shows the variance of the sharpness value (*d*-value) of one column line in this image and the curve approximated using a Gaussian function. We recognize that the POF is placed at the position of the peak *d*-value in this graph and the DOF expresses in-focus area size in this image. Thereby, we estimate the depth by estimating the peak position and the standard deviation of the approximated Gaussian curve, which depends on depth. However, from only one non-frontal image, it is difficult to obtain information on the POF and the DOF which are different for each pixel. Therefore, our proposed method obtains this information from two POFs and DOFs which are arranged as shown in Fig.3 in this our proposal.

B. Depth estimation based on sharpness model

In the following, we explain the estimation method using blur model based on the sharpness curve at each pixel.

We can obtain two sharpness values (d_-, d_+) at one pixel from two images, owing to two different optical settings. Hence, these two d-values are expressed by each Gaussian curve, as follows:

$$d_- = d_o + d_p \quad \exp\left(-\frac{1}{2}\left(\frac{y-\overline{y_-}}{\sigma}\right)^2\right), \qquad d_+ = d_o + \alpha d_p \exp\left(-\frac{1}{2}\left(\frac{y-\overline{y_+}}{\beta\sigma}\right)^2\right). \tag{8}$$

In these expressions, $\overline{y_-}$ and $\overline{y_+}$ are the positions of POF, and d_o and d_p are the d-values for offset and peak. α is difference of incident light determined by the image sensor's sensitivity and aperture size. Moreover, standard deviations of curves are also different (σ and $\beta\sigma$), which result from differences due to a setting of the f-number. In this way, these coefficient (α, β), which are obtained easily from d-values distribution such as Fig.2(b), are very important to facilitate setting of optics and image sensor.

Herewith, the following expression can be obtained from (8) to remove d_p, which depends not on depth but rather on image contrast:

$$\frac{\alpha(d_- - d_o)}{d_+ - d_o} = \exp\left(\frac{(y-\overline{y_+})^2 - \beta^2(y-\overline{y_-})^2}{2\beta^2\sigma^2}\right). \tag{9}$$

Note that $\overline{y_-}, \overline{y_+}$ and σ can be expressed as inversely proportional to the depth, because the POF position and the DOF limits in real space are expressed by linear equations, such as (1) and (6). Additionally, if the values of $\overline{y_-}$ and $\overline{y_+}$ are known for each depth (for example, by capturing a pattern such as that shown in Fig.2), then σ for each depth can also be obtained by using the following expression:

$$\sigma = \sqrt{\frac{(y-\overline{y_+})^2 - \beta^2(y-\overline{y_-})^2}{2\beta^2 \log\left(\frac{\alpha(d_- - d_o)}{d_+ - d_o}\right)}}. \tag{10}$$

Therefore, because the relations between the three parameters ($\overline{y_-}$, $\overline{y_+}$ and σ) and the depth are known, the depth value is obtained by finding the depth that satisfies expression (9). Herein, we estimated the correct depth value by using the method of least squares on the block of $m \times n$ pixels, using the following expression:

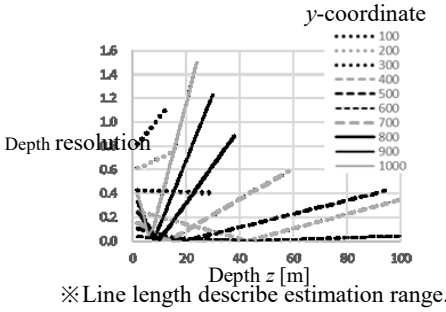
$$S = \sum_{i,k=1}^{m,n}\left(\log\frac{\alpha(d_{-(i,k)} - d_o)}{d_{+(i,k)} - d_o} - \left(\frac{(y_k-\overline{y_+})^2 - \beta^2(y_k-\overline{y_-})^2}{2\beta^2\sigma^2}\right)\right)^2. \tag{11}$$

Here, $d_{-(i,k)}$ and $d_{+(i,k)}$ are d-values that correspond to each coordinate on each $m \times n$ input block. In addition, this method is effective in restraining any noise in the d-values.

In the result from the above, It led that our method can allocate different range and different resolution at each y-coordinate as shown in the graph of Fig.4. Hence, the center area of an image which include the vanishing point can have wide estimation range. On the other hand, the upper and lower areas can have high estimation resolution (Fig4). It is useful for the advantage in the many situations effectively.

C. Depth estimation using neural network

However, it causes some estimation error based on the difference between the actual camera system and the optical



※Line length describe estimation range.

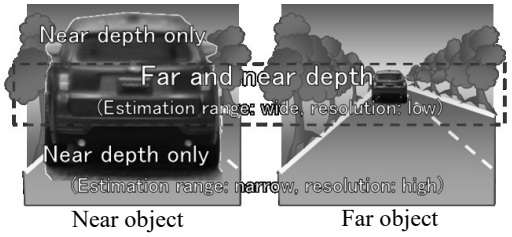


Fig 4. Estimation range and resolution on our method.

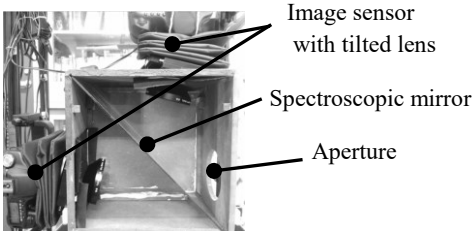


Fig.5 Our camera system with tilted optics.

theory based on blur model. Our method uses positively the circumference of the lens by using tilted optics which has wide angle. Hence, it is easy for our method to be affected by the influence of the optical aberration. Therefore, real sharpness model which is approximated by the Gaussian function is different slightly from ideal optical theory. Especially, depth errors at the top area and the bottom area on the image tend to increase.

In order to reduce the error, we propose to adopt the neural network to calculate the depth value from d-values and y-coordinate. The neural network which our use is general feedforward type consisted of full-connected layer only. Input layer has three units which receive two d-values and y-coordinate at each pixel, and output layer has one unit which puts one depth value. Each hidden layer has the rectified linear function as the activation function. The output layer only uses the identical mapping as activate function.

IV. EXPERIMENT

In this chapter, we explain our experiment details for performance of our method.

A. Our experiment environment

To test our proposed method, we performed experiments by using the camera system with a spectroscopic mirror (Fig.5). Hence, we simultaneously obtained two different blurred images through each tilted lens. These images were 8-bit monochrome.

Our camera system was set up using the conditions described by (1). From the relationship depth and peak position of *d*-values, we found that the two non-vertical POFs intersect at about 27.0m (Fig.3). In this way, we set up that the lens was tilted at angles $\theta \approx \pm 15°$. In this case, the theoretical value of L_z in (1) and the POF angle φ were 93mm and $\mp 0.7°$ respectively; the real POF angle appeared nearly horizontal. In addition, we used a lens for which the focal length was $f = 90\text{mm}$ and both aperture ratio was almost 5.6 (which could not be able to set exactly in our device). The sizes of both images were 1056×1584 pixels.

The neural network is constructed by the major deep learning library which is Keras based on Google Tensorflow. The network has 13 hidden layers which include 510 hidden units. In the learning process, the loss function is mean-absolute-error, the optimizer algorism of the learning rate is Adam, the amount of epochs is 60, the batch size is 4000. Herein, we captured two images by each depth position between about 10m to 50m; total is 16 images. As the result, 918410 of the learning data sets prepared; thereby, these data set are limited by threshold *d*-value larger than 15. Furthermore, these input data sets are normalized by average subtraction and deviation division.

B. Experimental results

In what follows, we described the results of using our proposed method with the above experiment environment.

In the estimation process, the input image which was taken by the 8-bit monochrome image like the learning process is shown in Fig.6. We set the target object on the floor which placed at 17.1 m which is the ground truth. The amount of the detectable pixels which are larger than threshold of *d*-value is 5046; the threshold of *d*-value is equals to that of learning process. Then, histogram of the depth value was generated for pixel by 1m as shown in Fig.7. By comparing the ground truth of depth obtained with laser measuring, we got the mean absolute error of the estimated depth based optical theory which is about 7.1m [9]. Then, we introduce the neural network for outputting depth value from inputting *d*-values and *y*-coordinate. As a result, we could improve the accuracy of our method; the mean absolute error is 2.1m. Hence, it is effective for our proposed method to introduce the neural network.

V. CONCLUSION

In this paper, we proposed to introduce the neural network to the depth estimation method with tilted optics. Additionally, we clarified that the performance of using the neural network is superior to the performance of using blur theoretical model directly. Hence, we succeeded in improving our method satisfies the requirements of these applications, such as real-time processing (only two images), wide detection area (passive sensing), sufficient image brightness (wide aperture), and robust use in automotive environments (small sensor interval).

In the future, we are planning to revise the estimation accuracy. In our result, we can slightly observe the influence of the overlearning, which occurs at the depth of 11m and 25m that is capturing point of learning data. In the real, if the learning steps (epochs) are increased extremely, these influence more increases. Hence, we are going to correct images at more depth positions to reduce the influence of overlearning.

Fig.6 Target image board value.

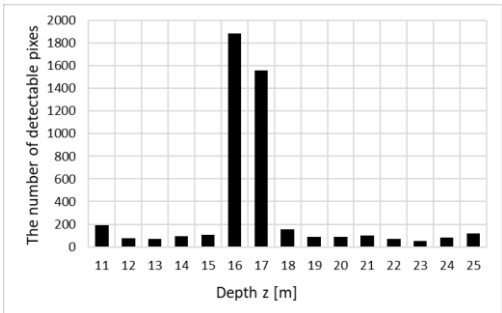


Fig.7 Histogram of depth value.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers 15K00365.

REFERENCES

[1] S.Tokoro, K.Kuroda, A.Kawakubo, K.Fujita, H.Fujinami, Electronically Scanned Millimeter-wave Radar for Pre-Crash Safety and Adaptive Cruise Control System, IEEE Intelligent Vehicles Symposium, pp.304-309, 2003.

[2] S.Ahn, J.Choi, N.L.Doh, W.K.Chung, A practical approach for ekf-slam in an indoor environment: fusing ultrasonic sensors and stereo camera, Auton.Robots, vol.24, no.3, pp.315-335, 2008.

[3] U.Franke, A.Joos, Real-time stereo Vision for Urban Traffic Scene Understanding, IEEE Intelligent Vehicles Symposium 2000, pp.273-278, 2000.

[4] M.Bertozzi, A.Broggi, A.Fascioli, S.Nichele, Stereo Vision-based Vehicle Detection, IEEE Intelligent Vehicles Symposium, IV, pp.39-44, 2000.

[5] T.B.Moeslund, C.B.Madsen, M.M.Trivedi, Improving stereo camera depth measurements and benefiting from intermediate results, IEEE Intelligent Vehicles Symposium, pp.935-940, 2010.

[6] V. Aslantas, D.T. Pham, Depth from automatic defocusing, OPTICS EXPRESS, Vol.15, No. 3, pp.1011-1023, 2007.

[7] H.Ikeoka, H.Kashiyama, T.Hamamoto, K.Kodama, Depth Estimation by Smart Imager Sensor Using Multiple Focus Images, The Institute of Image Information and Television Engineers, 62, 3, pp.384-391, 2008.

[8] H.Oku, M.Ishikawa, High-Speed Liquid Lens for Computer Vision, IEEE International Conference on Robotics and Automation Anchorage Convention District, pp.2643-2648, 2010.

[9] H. Ikeoka, T. Murata, M. Okuwaki, T. Hamamoto, DEPTH ESTIMATION FOR AUTOMOTIVE WITH TILTED OPTICS IMAGING, IEEE International Conference on Image Processing(ICIP), 5 pages, 2014.