PAPER *Special Issue on Computer Vision and Its Applications*

A Method for Depth Extraction by Motion Parallax*

Terunori MORI† *and* **Masanobu YAMAMOTO**†, ***Members***

SUMMARY A Dynamic Depth Extraction Method (DDEM) is proposed, which measures the time required for an edge to move through a known distance on the image plane and hence is able to calculate depth. Experimental results for three vertical bars in different depths show that the mean depths obtained by DDEM were almost the same as those obtained from direct measurement. The fluctuation of obtained depth was about 3.6%, which corresponds to one half frame difference in matching time of the near bar. Three kinds of thresholds (λ_1, λ_2 and λ_3) were introduced to reduce the noise affection. There was a wide range of thresholds for which the depth can be extracted stably. The DDEM was also successfully applied to recovering 3D structure of a complicated room.

1. Introduction

There are many depth extraction systems which are employed in the human visual information processing system such as binocular disparity, binocular convergence, gradient of texture and motion parallax. We propose a depth extraction method related to motion parallax. In this paper we describe a dynamic depth extraction method (DDEM) which tracks the time taken for an object to move through a known distance on the image plane. Experimental results are also reported.

Motion parallax and image sequence analysis have been used by many researchers. Our approach (DDEM) has following advantages.

(a) The Epipolar-Plane Image Analysis proposed by Bolles et al. (1987)[1] and Yamamoto (1986)[2] requires a huge amount of memory and involves additive picture processing such as straight line extraction on the epipolar-plane image. The DDEM method is able to process a image sequence without requiring large amount of memory nor complex image processing.

(b) The existing stereo method of searching for corresponding points and extracting disparity based on correlation can involve enormous calculation. Matthies et al. (1988)[3] used motion parallax and image sequences for depth estimation. In this work the search of corresponding points is comparatively easy, because the camera motion between frames is small. However, the distance moved on the image plane between frames is not accurate and so some complex calculation and integration of global disparity using methods such as the Kalman filter and regularization are required for determining accurate disparity. These problems arise because the number of pixels moved on the degitized image plane during a constant time interval is counted. The DDEM method does not require such complex processing because the method can calculate the time taken for an object to move through a known distance.

(c) The trade-off between the occlusion and the correspondence problem on the one hand and the accuracy of depth extraction on the other hand is involved in the traditional methods, because the accuracy of depth depends on the distance between the corresponding points there. The accuracy of depth in the DDEM depends principally on the number of frames per second, rather than the image resolution or the total distance of camera motion. Therefore, such a problem does not appear seriously in the DDEM method as in Bolles[1], Yamamoto[2] and Matthies et al.[3]

(d) In principle, the DDEM method is applicable to arbitrary camera attitudes.

2. Velocity Representation

When the observer (camera) moves in a perpendicular direction to the visual line (the optical axis), objects which are located close to the camera appear to move quickly while objects further away move more slowly. Therefore, if the velocity of an object is observed and the velocity of the observer is known, the depth of the object can be obtained.

Some examples of velocity representation[4] are as follows:

$$v = dx/dt \qquad (1)^{(5)}$$

$$v = -[dI/dt]/[dI/dx] \qquad (2)^{(6)-(8)}$$

Manuscript received January 9, 1991.
Manuscript revised May 18, 1991.

† The authors are with Electrotechnical Laboratory, Tsukuba-shi, 305 Japan.

* Based on "A dynamic depth extraction method" by T. Mori and M. Yamamoto which appeared in the Third International Conference on Computer Vision; Osaka, December 4-7, 1990; pp. 672-676. © 1990 IEEE.

$$v = W_t / W_x \qquad (3)^{(9)}$$

where $I(x, y)$ is the brightness at an image point (x, y), W_t the temporal frequency and W_x the spatial frequency.

There are two approaches to obtaining a solution in representation (1) above.

(a) Measurement of the distance which an edge (object) moves during the known time interval[(1)-(3),(10)].

(b) Measurement of the time required for an edge to move through a fixed distance

In the first case, the situation is similar to the measurement of binocular disparity, and the problems of stereo correspondence and occlusion will occur if the observer moves an appreciable distance. In the second case, it is possible to use a large number of sequential images instead of the single binocular stereo pair. Therefore, a large camera motion is not necessary to obtain an accurate depth, and the above correspondence and occlusion problems will seldom occur when the fixed distance interval is small (e.g., one pixel). Even in the case of occlusion, we can still obtain the time when the occlusion occurs using the measured time of the nearest object.

3. Dynamic Depth Extraction Method

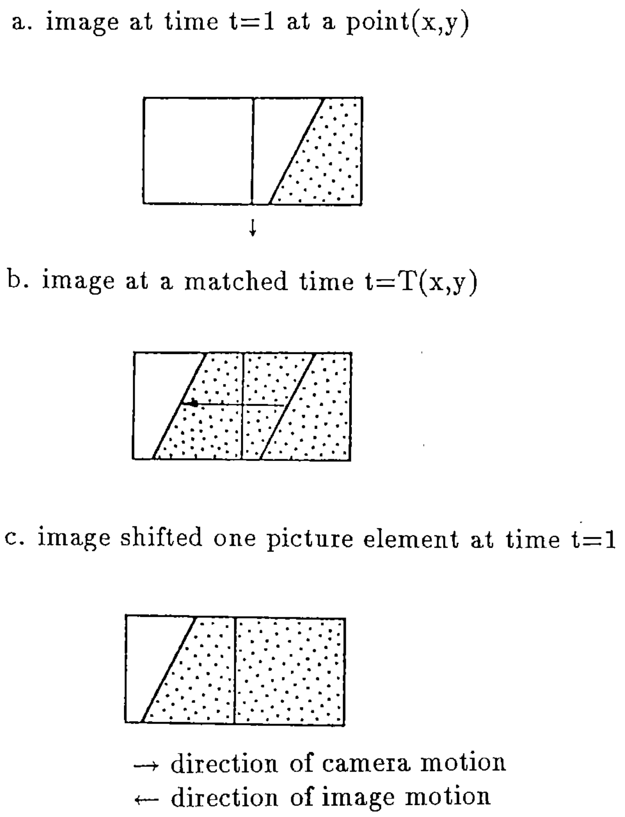
The DDEM consists of the extraction of the time required for an object to move through a fixed distance of one pixel and the calculation of the depth value at each pixel.

The objects we consider are the edges which are identified by examining the brightness gradient. The time taken for a particular edge to move through a fixed distance of one pixel is computed for all edges. This time is computed by comparing the image at any time t with the target image which is generated from the initial image by shifting it with one pixel. The method calculates the Degree of Mismatching for all pixels and stores the result in a point structure which monitors the current state of processing. A total of 128 images are acquired over the observation period in our experiments.

3.1 Extracting the Time $T(x, y)$ Required for an Object to Move a Fixed Distance

$I(x, y, t)$ is the brightness at a location (x, y) on the image plane at time t recorded by a TV camera moving from left to right through a fixed interval of 0.3 mm, where $(1 \leqq x, y, t \leqq 128)$. $G(x, y, t)$ is the gradient of brightness in the horizontal direction, and is defined as

$$G(x, y, t) = I(x+1, y, t) - I(x, y, t).$$



a. image at time t=1 at a point(x,y)

b. image at a matched time t=T(x,y)

c. image shifted one picture element at time t=1

Fig. 1 Principle of extracting the moving time $T(x, y)$.

$I_0(x, y)$ is the target image plane produced from the first image $I(x, y, 1)$ by shifting with one picture element to the left. $G_0(x, y)$ is the associated brightness gradient (See Fig. 1c). These are defined as

$$I_0(x, y) = I(x+1, y, 1),$$

$$G_0(x, y) = I(x+2, y, 1) - I(x+1, y, 1),$$

We introduce $S(x, y)$ as a plane which monitors the current state of processing. The initial values of $S(x, y)$ are defined as

$$S(x, y) = \begin{cases} 1 \text{ if } G_0(x, y) \geqq \lambda_1 > 0 \\ 0 \text{ if } G_0(x, y) < \lambda_1 \end{cases}$$

where λ_1 is a threshold value which is used to detect edge points. We wish to extract the time $T(x, y)$ for each point, when the following condition is satisfied (See Fig. 1b),

$$\begin{aligned} I_0(x, y) &= I(x, y, T(x, y)), \\ G_0(x, y) &= G(x, y, T(x, y)). \end{aligned} \qquad (4)$$

Unfortunately this condition (4) is seldom completely satisfied. Therefore, we extract $T(x, y)$ as a mean value of the time when the condition (4) is approximately satisfied. We also introduce $d'(x, y)$ as a plane which stores a current weighting value. The initial value of $d'(x, y)$ and $T(x, y)$ are zero. The Degree of Mismatching is defined as follows,

$$d(x, y, t) = |I(x, y, t) - I_0(x, y)| + |G(x, y, t)$$

$$- G_0(x, y)| \qquad (5)$$

This quantity $d(x, y, t)$ is introduced to judge whether or not an edge moved about one pixel on the image plane. When an edge moved just one pixel at time t, $d(x, y, t)$ must equal zero, because $I(x, y, t)$ equals $I_0(x, y)$ and $G(x, y, t)$ equals $G_0(x, y)$.

The following calculations are carried out for all points where $S(x, y) = 1$, until all values of $S(x, y)$ are changed to 0 or until the end of image plane ($t = 128$) is reached.

(i) if $d(x, y, t) > \lambda_2 (> \lambda_3)$ and $T(x, y) \neq 0$ then

$S(x, y) = 0$

(ii) if $d(x, y, t) < \lambda_3$ then

$T(x, y) = [d'(x, y) * T(x, y) + (\lambda_3 - d(x, y, t)) * t]/[d'(x, y) + (\lambda_3 - d(x, y, t))]$

$d'(x, y) = d'(x, y) + (\lambda_3 - d(x, y, t))$

λ_3 is a threshold to judge whether the Degree of Mismatching d is small enough or not, and λ_2 a threshold which is used to eliminate false edges. Figure 2 shows the change of $d(x, y, t)$ with time t and the extracted moving time $T(x, y)$ at a location (x, y).

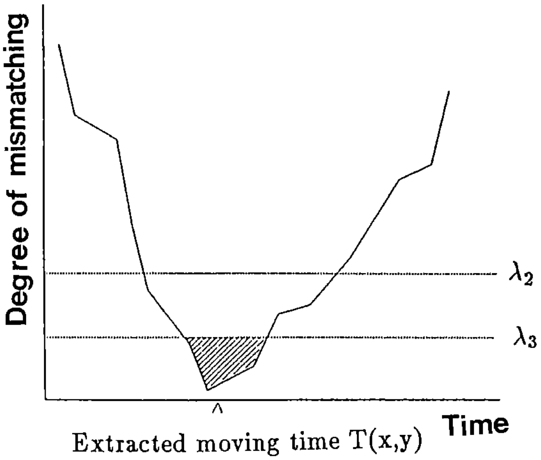


Fig. 2 Change of degree of mismatching with time.

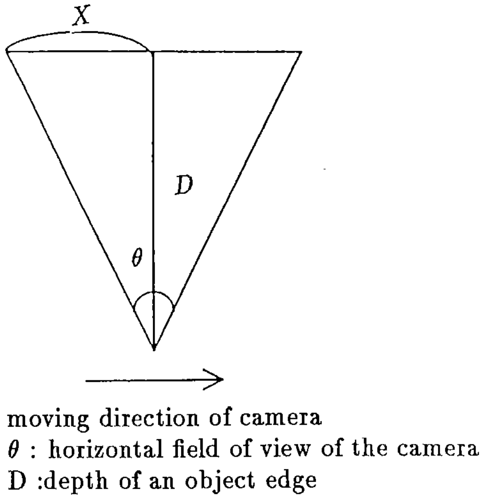

X
D
θ
moving direction of camera
θ : horizontal field of view of the camera
D :depth of an object edge

Fig. 3 Relation between variables D and θ.

3.2 Calculation of Depth $D(x, y)$ from $T(x, y)$

Let θ(23.55°) be the size of the horizontal visual field of camera, $D(x, y)$ be the depth of point (x, y), X be a half of the horizontal visual field width at $D(x, y)$ and δ(0.3 mm) be the moving pitch of the camera as shown in Fig. 3. Then, the following equation is obtained,

$$\tan(\theta/2) = X/D(x, y) \qquad (6)$$

The camera must move $X/64$ for the edge to move one pixel as the image has the size of 128×128 pixels and the moving distance in $(T(x, y) - 1)$ is $\delta * (T(x, y) - 1)$. Hence

$$X/64 = \delta * (T(x, y) - 1) \qquad (7)$$

From the Eqs. (6) and (7), the depth $D(x, y)$ is obtained as follows,

$$D(x, y) = 64 * \delta * (T(x, y) - 1)/\tan(\theta/2) \qquad (8)$$

4. Experimental Results

4.1 Three Vertical Bars

We applied the depth extraction algorithm to a series of scenes consisting of three vertical bars in different depths. The brightness of the scene was quantized to

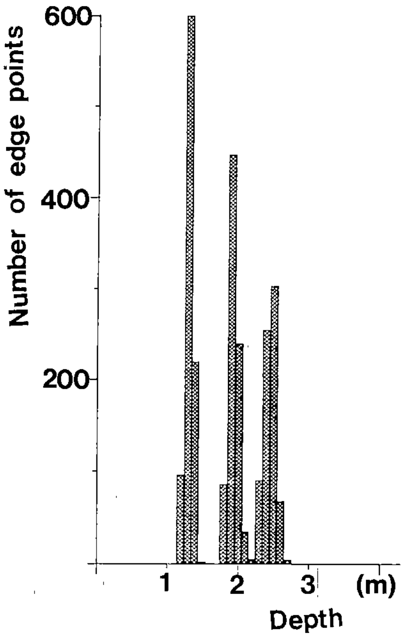


Fig. 4 Histogram of points with depth.

Table 1 Mean depth of bars.

	Near bar	Middle bar	Far bar
Mean depth D (mm)	1363	1978	2499
Standard Deviation S.D.(mm)	51	69	88
S.D. / D	0.0374	0.0351	0.0353
Measured distance (mm)	1360±5	1980±5	2500±5

256 levels. The maximum brightness of the scene was 120 and that of the background was about 20.

The histogram of computed depths on the three vertical bars is shown in Fig. 4. Table 1 shows the mean distance (D), the standard deviation ($S.D.$) and the relative standard deviation ($S.D./\,D$). The obtained mean distances were almost the same as those obtained from direct measurement. The $S.\,D.$ of the near bar corresponds to one frame difference in matching time.

Figure 5 shows the change of error in the obtained distance and the number of edge points for the thresholds λ_1, $(\lambda_2-\lambda_3)$ and λ_3, respectively.

Figure 6 shows the change of $S.D./\,D$ with the threshold λ_1, $(\lambda_2-\lambda_3)$, λ_3 respectively. These results show that changing the thresholds do not appreciably affect the computed depth, i.e., there is a wide range of thresholds for which the depth can be extracted stably.

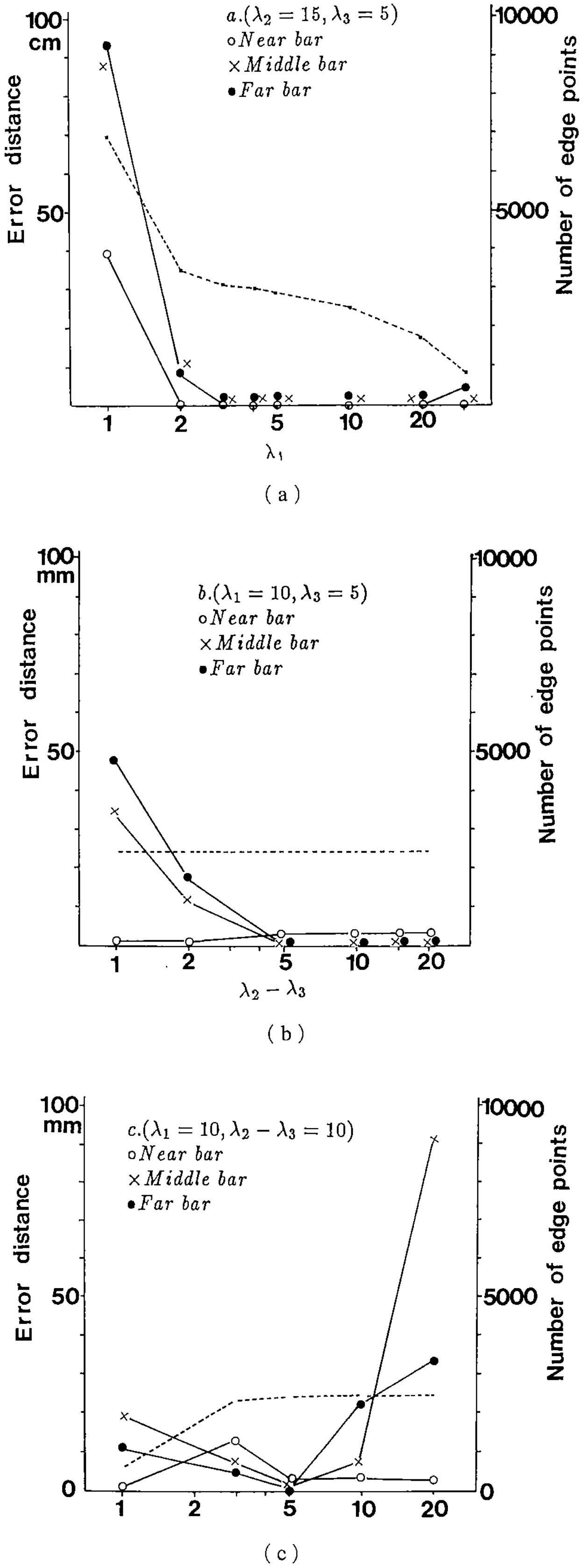
4. 2 Complicated Room Scene

As an example of the real world, we applied the algorithm to a series of scenes of a fairly complicated room (See Fig. 7). The result is shown in Fig. 8.

The values in Fig. 8 show the obtained depth. The unit of the value is 10 cm. The obtained values were approximately equal to the real distance.

4. 3 Discussion

The region where the DDEM can extract by 128 frames, which were observed at a moving pitch of 0.3 mm, is about 0.1-12 m. As each image was observed at a rate of 33 ms per frame, the velocity of camera motion was about 0.032 km/h. The total moving time and the total moving distance of camera were about 4.3 seconds and about 4 cm, respectively. The turn-around time for 128 frames was about 4 seconds by FACOM computer M380. The camera velocity was low and the turn-around time was a little long in the current experimental conditions. However, assuming



(a)

(b)

(c)

Fig. 5 Change of the error distance and the number of edge points with λ_1, λ_2 and λ_3.

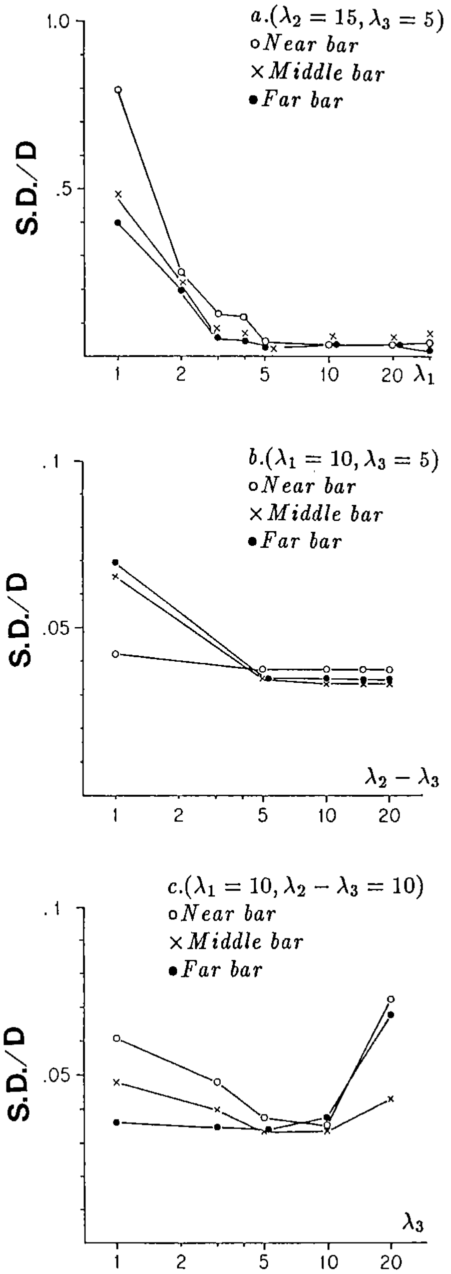


Fig. 6 Change of $S.\ D./D$ with λ_1, λ_2 and λ_3.

ten processors and a high speed camera of more than 300 frames/sec (presently available), we can use the camera velocity of 0.32 km/h (9 cm/sec) and the turn-around time to obtain a depth map of a region of 0.1 -12 m will be about 400 ms. Then, the DDEM is applicable to the vision system of a mobile robot.

5. Conclusions and Future Work

We have proposed a dynamic depth extraction method in which the time required for an edge to move through a fixed distance on the image plane is extracted at each point in parallel. The experimental results for three vertical bars in different depths show that the

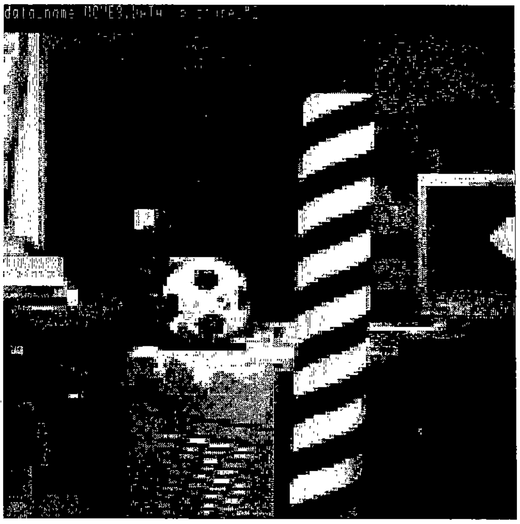


Fig. 7 Complicated room scene.

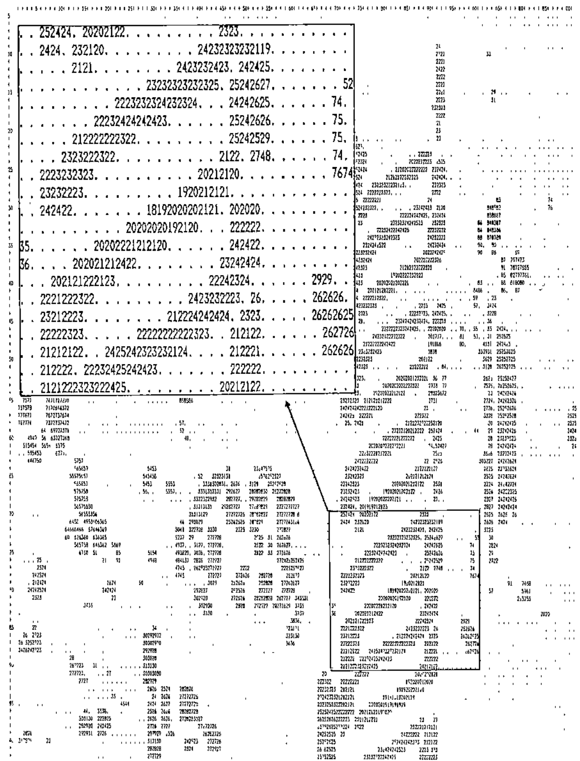


Fig. 8 Result for complicated room scene.

mean depths obtained by DDEM were almost the same as those obtained from direct measurement. The fluctuation of obtained depths was about 3.6%, which corresponds to one half frame difference in matching time of the near bar. Three kinds of thresholds (λ_1, λ_2 and λ_3) were introduced to reduce the noise affection. There was a wide range of thresholds for which the depth can be extracted stably. The DDEM was also successfully applied to a complicated room scene. The DDEM was applied at the point where there is a horizontal brightness gradient. Therefore, we need to combine this method with other procedures in order to obtain a more complete depth map which includes horizontal edges and other object points. In future work we intend to address the problem of occlusion which occurs when we wish to construct the 3D structure of a large region. As we can predict the time when a near object occludes a far object using the moving time of the near object, this problem can be solved.

Acknowledgements

The authors would like to thank Mr. S. Morioka and Mr. H. Endou for writing the simulation programs and Dr. R. Baldwin for reading the munuscript.

Note: The preliminary version of this paper was submitted to the Japanese pattent[11] and Trans. IEICE[12].

References

(1) Bolles R. C., Baker H. H. and Marimont D. H.: "Epipolar-plane image analysis: An approach to determining structure from motion", International Journal of Computer Vision, 1, pp. 7-55 (1987).

(2) Yamamoto M.: "Determining three-dimensional structure from image sequences given by horizontal and vertical moving camera", Trans. IECE Japan, **J69-D**, pp. 1632-1638 (1986).

(3) Matthies L., Szeliski R. and Kanade T.: "Incremental estimation of depth maps from image sequences", Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 366-375 (1988).

(4) Nakayama K.: "Biological image motion processing: A review", Vision Research, 25, pp. 625-660 (1985).

(5) von Hassenstein B. and Reichart W.: "Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswer tung bei der Bewegungsperzption des Russelkafers Chlorophanus", Z. Naturforsch., 11b, pp. 513-524 (1956).

(6) Cafforio C. and Rocca F.: "Method for measuring small displacements of television images", IEEE Trans. Inf. Theory, **IT-22**, 5, pp. 573 (1976).

(7) Limb J. O. and Murphy J. A.: "Estimation of velocity of moving images in television signals", C. G. I. P., 4, pp. 311 (1975).

(8) Skifstad K. and Jain R.: "Range estimation from intensity gradient analysis", Mashine Vision and Applications, 2, pp. 81-102 (1989).

(9) ed. Levinsion Z.: "Image motion", J. O. S. A. (A2) (1985).

(10) Moravec H. P.: "Robot rover visual navigation", UMI Research Press (1981).

(11) Mori T. and Yamamoto M.: "Three dimensional structure extraction method", Japanese patent, S62-214482 (1987).

(12) Mori T.: "A dynamic depth extraction method", Trans. IEICE, **J73-D-II**, pp. 955-960 (1990).

Terunori Mori was born in Ichinomiya, Japan on June 15, 1943. He received B.S. degree in 1966 and M.S. degree in 1968 in Physics from The Nagoya University and Ph.D. degree in Electronic Engineering from The Osaka University in 1977. From 1971, he was a researcher at Electrotechnical Laboratory. He has been working on OCR, Motion Perception, Human Memory and Computer Vision. Currently, he is a senior researcher at Image Understanding Section of Electrotechnical Laboratory. He received 1973 Yonezawa Award from the IEICE.

Masanobu Yamamoto was born in Shimonoseki, Yamaguchi, Japan in 1951. He received the B.S. degree in control engineering from the Kyusyu Institute of Technology, Japan in 1973, and M.S. degree and Dr.Eng. degree from the Tokyo Institute of Technology in 1975 and 1988, respectively. Since 1975, he has worked in the Electrotechnical Laboratory of Ministry of International Trade and Industry, in Japan. He is a senior research scientist of the Computer Vision Section in this laboratory. His major research interests lie in motion analysis, computer vision, robotics, vision architecture and discovery system. Dr. Yamamoto was a recipient of the SIG Research Award in 1987 from The Information Processing Society of Japan.