# 6—2
# Toward Occlusion-Free Depth Estimation for Video Production

Jong-Il Park and Seiki Inoue

ATR Media Integration & Communications Research Labs.
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
Email: {pji,sinoue}@mic.atr.co.jp

**Abstract:** We present a method to estimate a dense and sharp depth map using multiple cameras for the application to flexible video production. A key issue for obtaining sharp depth map is how to overcome the harmful influence of occlusion. Thus, we first propose to selectively use the depth information from multiple cameras. With a simple sort and discard technique, we resolve the occlusion problem considerably at a slight sacrifice of noise tolerance. However, boundary overreach of more textured area to less textured area at object boundaries still remains to be solved. We observed that the amount of boundary overreach is less than half the size of the matching window and, unlike usual stereo matching, the boundary overreach with the proposed occlusion-overcoming method shows very abrupt transition. Based on these observations, we propose a hierarchical estimation scheme that attempts to reduce boundary overreach such that edges of the depth map coincide with object boundaries on the one hand, and to reduce noisy estimates due to insufficient size of matching window on the other hand. We show the hierarchical method can produce a sharp depth map for a variety of images.

## 1 Introduction

Recently, computer vision technology is widely used for achieving high degree of freedom and efficiency in video content creation. Thus, it is even said to be a kind of media technology [11].

We are developing a video component database in order to realize a flexible and versatile framework for video content creation [5]. It is based on the layered representation of video where a video sequence is regarded as a spatio-temporally ordered set of video components [15]. Video components are stored with various property information such as camera work, key words, depth, and the like in the database. We can freely select some video components from the database and enjoy arranging them in a spatio-temporal domain to make a new video and/or creating new video expressions by exploiting the given property information.

Among the information, one of the most important one would be depth. It is $Z$ value of the camera-centered coordinate of the corresponding object point for each pixel, where $Z$ axis is set to optical axis. Depth information corresponds to the spatial part in the spatio-temporal description of a scene. Thus, it takes a crucial role in making natural-looking videos and/or creating various video expressions with high degree of freedom using the video component database. Virtualized Reality [7], Z-keying for video composition [8], and 3D special video effects [13] are typical application using the depth information. Moreover, we can automatically generate multi-layer description of a scene using depth information [14].

For such application, dense and sharp depth map is strongly required. Here, "sharp" means that object boundary and/or depth discontinuity should be correct. In other words, correctness of depth map in shape is more important than precision of depth value. In this paper, how to get such dense and sharp depth map is the main theme. The proposed method is a novel hierarchical scheme combined with an occlusion-overcoming disparity estimator using stereo images from 5 cameras. Considering hardware feasibility, we confine the method to a signal-level processing.

## 2 Related Works

Stereo matching is a useful method in obtaining depth map from image. There are two approaches in stereo matching. One is area-based method and the other is feature-based method. Area-based method is, in general, used for obtaining a dense depth map [1]. However, it is well-known that the area-based stereo matching faces several problems such as lack of texture, occlusion, photometric change, repetitive pattern, and so on [1][2].

A considerable amount of effort has been exerted to cope with such problems in computer vision [2][10]. Almost methods to obtain dense depth map are computationally expensive or iterative. Among some exception is multiple-baseline stereo matching [16][12]. It demands more cameras but alleviates the problems of lack of texture and repetitive texture without much increase of computational complexity. Recently, a real-time depth mapper has been developed on the basis of multiple-baseline method [8].

However, little attention has been paid to clearing the occlusion problem in stereo matching. As Dhond and Aggarwal pointed out[2], the presence of occluding boundaries in the matching window tends to confuse the matcher and often giving an erroneous depth estimate. In fact, occlusion is one of the main culprits

to prevent from obtaining correct depth map in shape as we will show in this paper. In two-view stereo, occlusion problem is unavoidable and it is impossible to get correct match. Only some appropriate interpolation can fill such area based on some assumption and knowledge [1]. From the standpoint of correct match, multiple view(more than two) can give a clue to resolve the occlusion problem. When an area is occluded in an image from a camera, another camera located at a different position can see the area and give a correct match. Kanade *et al.*'s depth mapper does not seem to explicitly exploit this property although they touched the occlusion problem a little [8]. Recently, Nakamura *et al.* extensively studied the occlusion problem [9]. Using eye array camera, they analyze occlusion patterns quantitatively and propose a disparity estimation scheme which is capable of detecting occlusion, selecting a proper mask for correct match, and thus preventing from mismatch. However, it demands at least 9 cameras and furthermore, it does not provide a strategy for controlling the effect of matching-window size.

A very important issue underlying the area-based matching is the size of matching window [6][10]. It should be large enough to include enough intensity variation for characterizing an area. But it should be small enough to avoid projective distortion. Toward resolving such a dilemma, two approaches have been proposed. One is to use a locally adaptive window [6]. It searches for a window that produces the estimate of disparity with the least uncertainty for each pixel of an image. Considerable improvement is obtained from the aspects of smooth surface and sharp disparity edges. However, the problem is that it is iterative. The other approach is to use hierarchical coarse-to-fine scheme [2][10]. Traditional hierarchical schemes in general restrict searching range to within some neighborhood of the estimate of coarse resolution [3][4]. The concept is based on the assumption that the disparities of geometrically neighboring pixels are not much different. This works well if there is no discontinuity of the disparity to be estimated. However, disparity, the parameter we are to estimate, has many discontinuities, mainly at object boundaries.

Exact profile of the disparity around object boundaries is very important for the application of our concern such as video composition, video effects, scene description, and so on. Thus, we propose a novel hierarchical scheme combined with occlusion-overcoming strategy.

## 3 Depth Estimation

### 3.1 Configuration of 5 Camera System

The proposed configuration of multiple camera system is shown in Fig. 1. We put a camera at the center and a total of 4 cameras of the same specification to each direction of upper, lower, right, and left, separated by the same distance $L$.

Under the projection geometry of Fig. 1, an object point $P = (X, Y, Z)$ is projected to the point $p_0 = (x_0, y_0)$ on the image plane of the center(=base) camera, where $x_0 = F\frac{X}{Z}$ and $y_0 = F\frac{Y}{Z}$, and $F$ is the focal length. It is also projected to $p_i = (x_i, y_i)$ on image

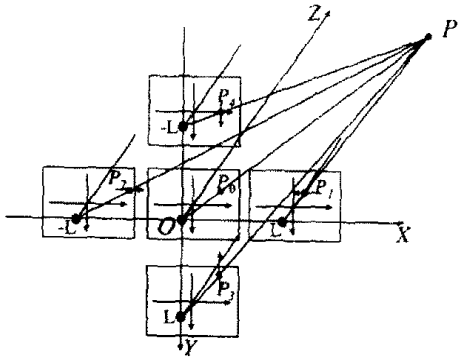


Figure 1: Projection geometry of camera system.

plane of each inspection camera $C_i$, $i = 1, 2, 3, 4$, where

$$x_i = F\frac{X - D_{i,x}}{Z}, \quad y_i = F\frac{Y - D_{i,y}}{Z}. \tag{1}$$

Here, the baseline stretch $D_i = (D_{i,x} D_{i,y})$'s are $D_1 = (L, 0)$, $D_2 = (-L, 0)$, $D_3 = (0, L)$, and $D_4 = (0, -L)$. In the configuration, the true disparity $d_t$ of the object point $P$ is $d_t = \frac{FL}{Z} = |p_i - p_0|$, for all $i$. Thus, by estimating the disparity, we can obtain the depth of an object point.

The camera configuration is based on the assumption that when a camera cannot give a correct match for a pixel because of occlusion, another camera located at the other side can give a good one. This holds good for almost occluding cases. In this sense, we may say, the more the number of cameras, the more chance to obtain a good match we have. However, if we consider practical implementation, the number of cameras should be small enough. Thus, we set the number of cameras to 5 which we think reasonable and seem to be the least number for coping with the occlusion problem without suffering from too much complexity of computation and calibration setup.

Optical axes of the 5 cameras are parallel and the cameras are synchronized. What we are to acquire is the depth map of the image from the center camera. Other cameras work as sensors in this sense. We now explain how we reduce possible bad matches and obtain a good one around occlusion area in the followings.

### 3.2 Occlusion-Overcoming Stereo Match

We use the sum of squared-difference as a matching measure. At a point $\mathbf{x}$ on the image plane of the base camera, the matching measure is calculated at each displacement $d$ for each camera $C_i$ by

$$e_i(\mathbf{x}, d) = \sum_{\mathbf{b} \in W} [I_0(\mathbf{x} + \mathbf{b}) - I_i(\mathbf{x} + \mathbf{b} + \mathbf{d}_i)]^2 \tag{2}$$

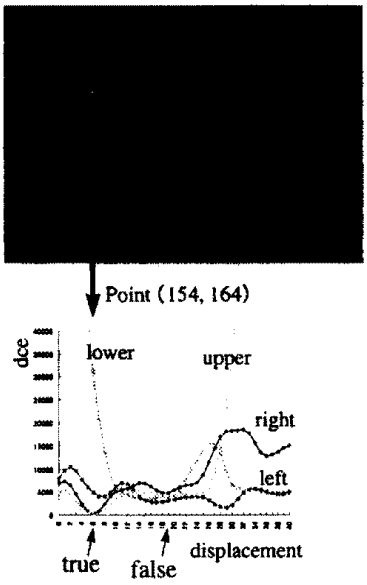


Figure 2: Illustration of the influence of occlusion in stereo matching from multiple cameras.

where $I_i$ is the intensity and $W$ is a matching window. The disparity should be the same for all cameras in the camera geometry such that $d = |\mathbf{d}_i|$ for all $i$.

A straightforward implementation of multiple-baseline stereo [8] using the camera configuration in Fig. 1 would be

$$\hat{d}(\mathbf{x}) = arg\min_{d} \sum_{i=1}^{4} e_i(\mathbf{x}, d). \quad (3)$$

It gives a good result if there is no discontinuity of depth. However, when there is a discontinuity of depth near the matching window for a pixel, we cannot expect all of the matching data from the 4 directions gives us useful information. On the contrary, some data, especially from the direction of occluded area, affect the estimation harmfully, which should be eliminated for a good estimation. We illustrate this phenomenon in Fig. 2 where the $e_i$ curves for a typical point around object boundary are shown (the size of matching window = 7x7 [pixels]). Due to the influence of occlusion, that is, the bad observation data from lower and left cameras, the matching tends to produce undesirable result($\hat{d} \approx 19$) where the true disparity is $d_t \approx 6$. We see why the matching based on eq.(3) cannot be successful in such area.

If we can eliminate such bad observations during the matching, a considerable improvement can be expected. Thus, we devise a simple sort and discard method based on the observation from $e_i$ curves.

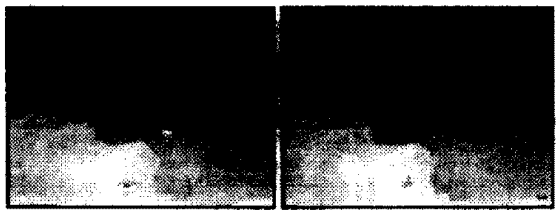

Figure 3: Depth map from multi-camera matching. Left map is obtained without occlusion-overcoming strategy and right map with the strategy. Matching window size is 15x15.

We assume that at least two data sets among the four are not corrupted by occlusion. At each displacement, we sort the difference $e_i(\mathbf{x}, d)$ and discard the largest 2 data among the given 4 data. We sum the two data that are considered as useful for the estimation of disparity. By repeating the collection and summation of data along the epipolar line, we get a 1-D curve for the estimation and consider the displacement which gives the minimum of the curve as the disparity of the pixel. In short, the estimation scheme can be described by

$$\hat{d}(\mathbf{x}) = arg\min_{d} \sum_{i=1}^{2} \tilde{e}_i(\mathbf{x}, d), \quad (4)$$

where $\tilde{e}_i$ is the sorted one of $e_i$ such that $\tilde{e}_i \leq \tilde{e}_j$ for all $i < j$.

As we see in the Fig 3, a considerable improvement is achieved around depth discontinuity by the above scheme at the slight sacrifice of noise tolerance as is expected [12]. The loss of noise tolerance will be compensated for by a hierarchical scheme in the following subsection. In the depth maps, we can observe two kinds of distortion. One is from occlusion. We see many noisy estimates around object boundaries in the left depth map while no such estimates in the right depth map. The other is *boundary overreach*. As Cochran *et al.* pointed out, the more strongly textured surface tends to leak into the less textured region. We see the disparity of higher texture tends to reach over the true edge of disparity and out to lower texture area in both of the depth map. The difference is that the depth map by the proposed occlusion-overcoming strategy shows clear and abrupt change around discontinuity. Moreover, the amount of the boundary overreach is roughly half of the size of the matching window as we see in the magnified depth map with edge map overwritten in Fig. 4. The size of matching window is 15x15 [pixels] and the size of the leaks is less than 7 pixels. This observation gives us an idea to implement a novel hierarchical scheme.

Smaller matching window is favorable in the aspect of reducing boundary overreach. But, smaller matching window gives us less reliable results of estimation.

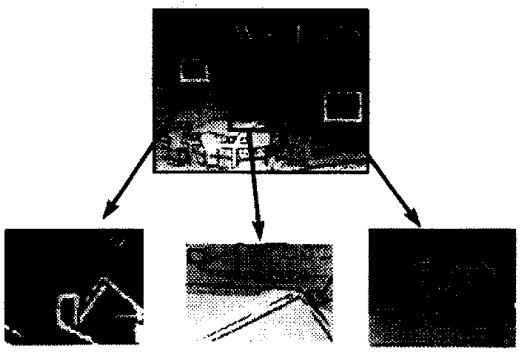

Figure 4: Boundary overreach. The amount of boundary overreach is about half the size of the matching window.

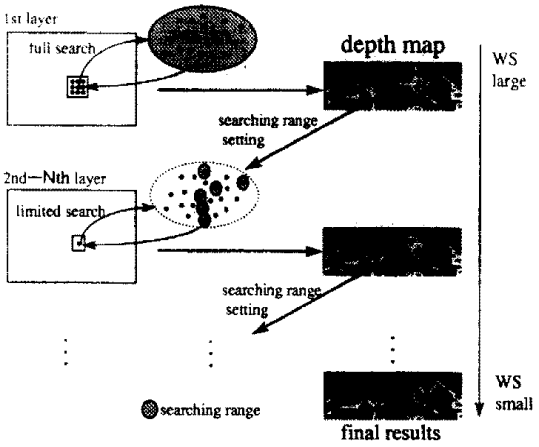


Figure 5: Illustration of the hierarchical scheme. Searching range is restricted to the estimates in the matching window of previous(=large matching window) layer except for the 1st layer.

There is a trade-off between reliability and geometric correctness of estimation with respect to the size of matching window. Thus, we propose a hierarchical scheme in order to cope with the trade-off problem as follows.

### 3.3 Hierarchical Estimation Scheme

The proposed hierarchical method is based on the observation that, when we use the occlusion-overcoming strategy, the correct disparity for boundary overreach area exists near(within half of the size of the matching window) the point in the disparity map in most cases. Thus, we first obtain a depth map with a large matching window. Assuming all the necessary depth values are included in the depth map, we successively refine the depth map with decreasing the size of matching window.

Figure 5 illustrates the concept of the proposed hierarchical method. First, we set the size of matching window(WS) to a large value. Then, we estimate the disparity of each pixel to a pixel accuracy based on the collection and summation of difference data over all searching range(eq.(4)) and we get the disparity map of the 1st layer. We assume the true disparity value of a pixel exists within the matching window of the pixel in the disparity map of the 1st layer.

Now, we reduce the size of the matching window by half. Then, the disparity is estimated by the same way(eq.(4)) except that the searching range is restricted to a set consisting of the disparity values of the upper layer within the window of the upper layer at the position. This restriction of searching range is based on the above observation about boundary overreach.

The procedure is repeated until the last layer where the size of matching window is 3x3.

When there is no disparity discontinuity around a point(within the matching window of the point), we don't need to estimate the disparity of the points again in the next layer. Instead, we just enhance the resolution of the disparity value to sub-pixel accuracy by quadratic fitting and no more update in the successive layers.

Figure 6: Depth map obtained by the hierarchical method without post-processing [left]. Edge map is overwritten [right].

We see very clear boundaries in the depth map of Fig. 6. It is obtained by using 3-layer(15x15 to 7x7 to 3x3) hierarchy. However, some noisy estimates can still be observed in the depth map. The noise can be efficiently eliminated by an edge-preserving post-processing. We use an adaptive order-statistics filter of Yang *et al.*[17] which shows excellent performance through our experiments. Figure 7 shows a result of post-processing. We see in the depth map that almost noisy estimates are eliminated while very sharp edges at the correct positions are preserved, which can be seen more clearly in Fig. 8.

Unlike other hierarchical coarse-to-fine strategies [10], the proposed method is fine-to-fine approach. It means we sacrifice the computational benefit of hierarchical approach. The proposed method can be easi-

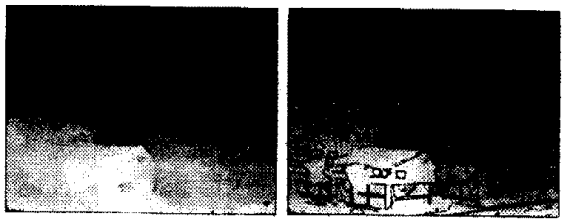

Figure 7: Depth map obtained by the hierarchical method with post-processing [left]. Edge map is overwritten [right]. A simple adaptive order-statistics filter is used for the post-processing.

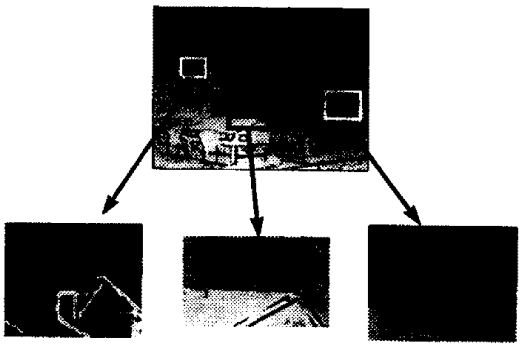

Figure 8: Magnified depth maps showing how well the proposed method can overcome the occlusion problem and the boundary overreach.

ly changed to coarse-to-fine method to get the benefit of computational saving at the sacrifice of correctness which would not be covered in this paper.

## 4 Experimental Results

We have tested the proposed algorithm for a variety of indoor images with 640[pixels]x480[lines] and 8-bit gray-scale resolution. They include some eye-array images in the image database of Tsukuba University and some indoor images shot in our laboratory.

### 4.1 Performance

We show some of the results obtained by the proposed hierarchical method in Fig. 9.. The left one is the depth map of "Santa" image in the database of Tsukuba Univ. Searching range $SR$ of the 1st layer is set to 50 pixel and 4-layer hierarchy(31x31 to 15x15 to 7x7 to 3x3) is used. We see the object edges exactly coincide with depth edges and the surfaces of depth map are very smooth. In the no texture areas, for example, the

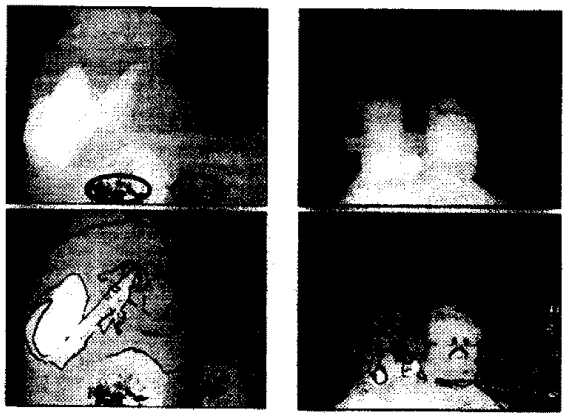

Figure 9: Depth map of "Santa" image[left] and "Lab" image[right]. Edges are overwritten in the lower maps.

foot of the stuff (circled in the figure), we see undesirable errors, which is the fundamental limitation of the area-based matching. The right one is the depth map of "Lab" image shot by SONY 3CCD DXC-930 cameras in our laboratory. The conditions are the same as the previous one. We can confirm the proposed method produces a sharp and correct depth map. Some errors around the stuffs seem to be due to lack of texture and too complex occlusion pattern (circled area 1) and saturation of intensity value(circled area 2).

Through many experiments, we can confirm that the proposed method works very well for a variety of scenes. It improves not only the correctness of depth map around object boundaries but also the smoothness of surface.

### 4.2 Application

A variety of video expression can be created by using a depth map of an image [14]. Figure 10 shows an example of video composition. We see the two video are merged in a natural way based on the depth information and the object boundaries are very sharp and correct. More application examples such as new view generation, automatic scene description, and special effects can be found in [14]

## 5 Concluding Remarks

We have presented a method to obtain a sharp and dense depth map based on a simple discard and summation of disparity information from 5 cameras implemented on a novel hierarchical estimation scheme. By using occlusion-overcoming strategy, we have reduced the harmful effects of occlusion considerably. Furthermore, based on the unique property of boundary overreach of the strategy, we have constructed a hierarchical estimator. Thus, we could achieve shape-correctness of depth map on the one hand, alleviate the problem of lack of texture on the other hand. The performance of the method has been verified by experimental results.

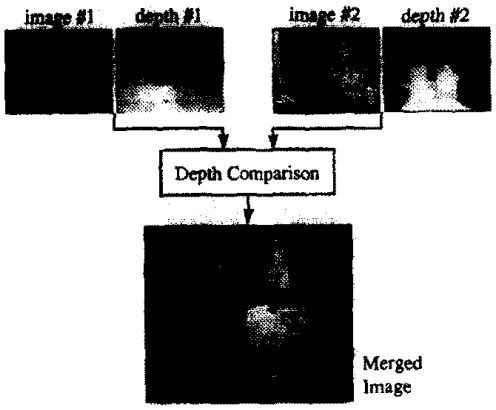


Figure 10: An example of video composition using the depth information.

With all the considerable improvement, the method sometimes fails to give a correct match for strongly concave areas, small holes, and narrow valley as can be reasonably predicted from the configuration of the camera system. The limitation does not seem to be resolved within the scope of non-iterative signal-level processing. Some high-level recognition and/or interactive scheme would be necessary for overcoming this limitation.

Since the obtainable performance of the estimation depends on images, it would not be unusual for the obtained one not to come up to the desired one. In such cases, some interactive interface should be prepared to fill the gap of the quality and thus to expand the scope of video expression. Thus, we are currently developing a user-friendly interface for post-processing of the estimation.

We have focused on only the spatial characteristics of a scene in this paper such that all of the processing is executed by the frame. Temporal property has not been considered. In order to obtain more satisfactory results, we may need to develop an integrated approach of spatial and temporal property of a scene. Therefore, a paradigm of motion and structure from multiple-baseline stereo would be very interesting as a future work.

## References

[1] S.D.Cochran and G.Medioni, "3-D surface description from binocular stereo," *IEEE Trans. PAMI*, vol.14, no.10, pp.981-994, Oct. 1992.

[2] U.Dhond and J.Aggarwal, "Structure from stereo: A review," *IEEE Trans. System, Man, and Cybernetics*, vol.19, no.6, pp.1489-1510, Nov./Dec. 1989.

[3] W.E.L.Grimson, "A computer implementation of a theory of human stereo vision," *Phil. Trans. Royal Soc. London*, vol.B292, pp.217-253, 1981.

[4] M.J.Hannah, "Bootstrap stereo," *Proc. ARPA Image Understanding Workshop*, pp.201-208, College Park, MD, Apr. 1980.

[5] S.Inoue, "Mental image expression by media integration - COMICS," *Proc. of 1st International Workshop on New Video Media Technology*, pp.47-52, Seoul, Korea, March 1996.

[6] T.Kanade and M.Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *IEEE Trans. PAMI*, vlo.16, no.9, pp.920-932, Sept. 1994.

[7] T.Kanade *et al.*, "Virtualized Reality: Concepts and early results," *Proc. IEEE Workshop on Representation of Visual Scenes*, pp.69-76, June 1995.

[8] T.Kanade *et al.*, "A stereo machine for video-rate dense depth mapping and its new applications," *Proc. IEEE CVPR'96*, pp.196-202, San Francisco, June 1996.

[9] Y.Nakamura *et al.*, "Occlusion detectable stereo - Occlusion patterns in camera matrix," *Proc. IEEE CVPR'96*, pp.371-378, San Francisco, June 1996.

[10] V.S.Nalwa, *A Guided Tour of Computer Vision*, Addison-Wesley, 1993.

[11] Y.Ohta, "Computer vision as media technology," *Proc. Image Sensing Symposium*, pp.265-270, 1996 (in Japanese).

[12] M.Okutomi and T.Kanade, "A multiple-baseline stereo," *IEEE Trans. PAMI*, vol.15, no.4, pp.353-363, April 1993.

[13] J.Park *et al.*, "Extraction of depth information for scene description and its application," *ITE'96*, pp.112-113, Nagoya, Japan, July 1996 (in Japanese).

[14] J.Park and S.Inoue, "Image expression based on disparity estimation from multiple cameras," *Proc. 3rd Joint Workshop on Multimedia Communications*, 7-1, Taegu, Korea, Oct. 1996.

[15] M.Shibata *et al.*, "Scene describing method for video production," *ITEJ Tech. Report*, vol.16, no.10, pp.19-24, Jan. 1992 (in Japanese).

[16] R.Tsai, "Multiframe image point matching and 3-D surface reconstruction," *IEEE Trans. PAMI*, vol.5, no.2, pp.159-174, March 1983.

[17] K.H. Yang, S.G. Lee, and C.W. Lee, "Image Restoration of Noisy Images Using OS Filters with Adaptive Windows," *J. Korean Insti. of Telematics and Electronics*, vol. 27, no. 1, pp.112-119, Jan. 1990 (in Korean).