Half-Diminished Reality Image Using Three RGB-D Sensors for Remote Control Robots

Kazuya Sugimoto*, Hiromitsu Fujii*, Atsushi Yamashita* and Hajime Asama*
* The University of Tokyo
Department of Precision Engineering
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
Email: sugimoto@robot.t.u-tokyo.ac.jp
fujii@robot.t.u-tokyo.ac.jp
yamashita@robot.t.u-tokyo.ac.jp
asama@robot.t.u-tokyo.ac.jp

***Abstract*—This paper presents a method to compose half-diminished reality images for remote control robots using three RGB-D sensors. Robots are designed to perform the work of humans during a disaster response to reduce secondary disasters. In these case, a operator controls the robot with the aid of images from a mounted camera. However, operating efficiency decreases significantly when the robot arm occludes target work objects in those images. Recently, the *Half-Diminished Reality* technique has been proposed in the field of computer vision. This technique is used for seeing through foreground objects and viewing occluded backgrounds. Accordingly, composing half-diminished reality images can enable the operator to see the target work objects occluded by the robot arm in only one image, which results in increased the operating efficiency.**

I. Introduction

There are many places considered to be too dangerous for humans to go during a disaster response when a natural calamity occurs and safety must be restored. For example, after the great eastern Japan earthquake and tsunami that occurred on March 11, 2011, buildings were destroyed and fires erupted, and a the high radiation area was caused by the accident at the Fukushima Daiichi Nuclear Plant. These incidents caused secondary disasters because the emergency responders faced significant dangers in these situations.

Thus, robots designed to work in place of human to reduce secondary disasters during disaster responses [1]. These robots are also called rescue robots, and many of them have been developed in recent years [2]. In particularly, there is a significant amount of research describing the development of remote control systems for managing tasks in disaster sites [3]. For example, robots with arms are used for tasks such as the operation of switching valves or debris removal in disaster sites as shown in Fig. 1. Operators control robots remotely with the aid of mounted camera images as shown in Fig. 1. However, the operating efficiency is significantly lower than by direct viewing [4]. This occurs because some target work objects are occluded by the robot arm in the camera image. Mounting multiple cameras is an effective technique to improve operating efficiency [4], [5]. Using multiple cameras enables to the operator to confirm the position and shape of the target work objects, because dead spaces that the operator cannot see are decreased. However, considerable skill and concentration are necessary for performing operations with multiple images, which may result in reduced operating efficiency. Thus, it is important to compose the image in a manner that enable the operator to see the target work objects occluded by the robot arm in only one image.

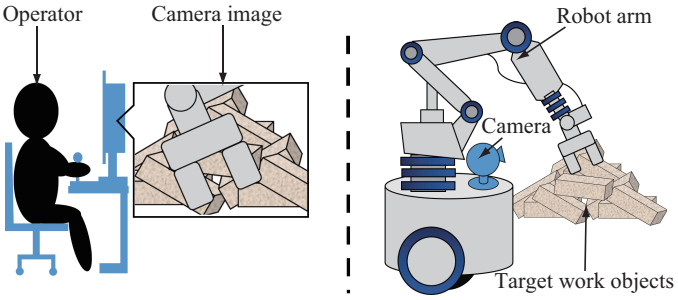


Fig. 1. Remote control system of a robot. Operators control robots remotely with the aid of a mounted camera. However, some target work objects are occluded by the robot arm in the camera image. Therefore, the operating efficiency decreases significantly.

Recently, in the field of computer vision, **Diminished Reality**（DR）has become a popular technique for removing obstacles in images and replacing them with an appropriate background image. DR has been actively researched in recent years [6], [7], [8]. In particular, **Half-Diminished Reality** is a technique for diminishing unwanted objects in images and, allowing the viewer to see the background. This technique is also called See-through [10], [13] or X-Ray [9], [11]. For example, this technique is used for diminishing walls of buildings in images, in order to view the background [9], [10], [11]. In addition, several studies of this technique have been also conducted for the entertainment [12] and broadcast industries [12].

Accordingly, composing half-diminished reality images enable to operator to view the target work objects occluded by the robot arm in a single one image. Thus, to increase operating efficiency, we present a method to compose half-diminished reality images for a remote control robot system.

II. Related Research

In most cases where half-diminished reality images are composed, the background occluded by an object in an image is obtained using one of the following two method types.

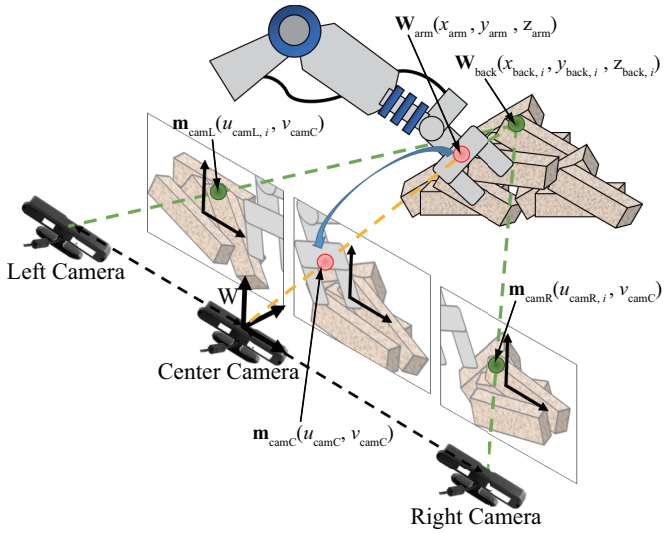


Fig. 2. Schematic view of experimental conditions. First, the arm area ($\mathbf{w_{arm}} = [x_{\mathrm{arm}}, y_{\mathrm{arm}}, z_{\mathrm{arm}}, 1]^T$) in the center camera image is calculated by equation(1). Second, the hidden area ($\mathbf{m}_{\mathrm{camL}} = [u_{\mathrm{camL},i}, v_{\mathrm{camL}}, f_{\mathrm{camL}}, 1]$ or $\mathbf{m}_{\mathrm{camR}} = [u_{\mathrm{camR},i}, v_{\mathrm{camR}}, f_{\mathrm{camR}}, 1]$) occluded by the arm area is estimated from the left camera image or the right camera image.

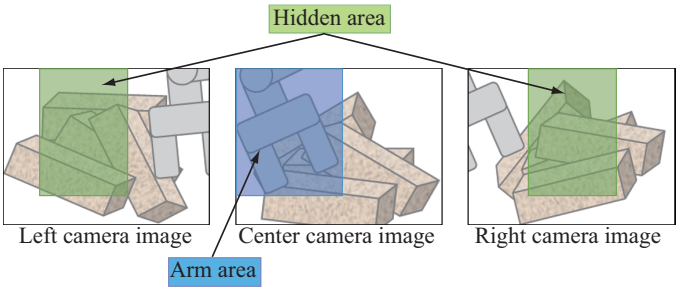


Fig. 3. Definition of each area in camera images. An area of the robot arm in the center camera is defined as an arm area, and areas occluded by the robot arm area in the other camera images are defined as hidden areas.

The first method type uses time series images [7], [8], [14]. In this method, the background is estimated using the time series image captured at the same position, in which the target work objects are not occluded. However, when operating a robot, the camera image changes from moment to moment because the robot moves and the environment changes dynamically. Thus, if time series images are used to compose the resulting image cannot be relied upon for operating the robot. Therefore, this method is not suitable for this research, because the proposed image is composed only of time series images viewed from the same position.

The second method uses multiple view images [6], [9], [10], [11], [12], [13]. In this method, the background is estimated using camera images viewed from other positions. Thus, this method can compose the proposed image independent of the robot's movement and changes in the environmental. In our research, we decide to use this multiple view-based method. However, in many research efforts the background was assumed to be a two-dimensional surface, because the

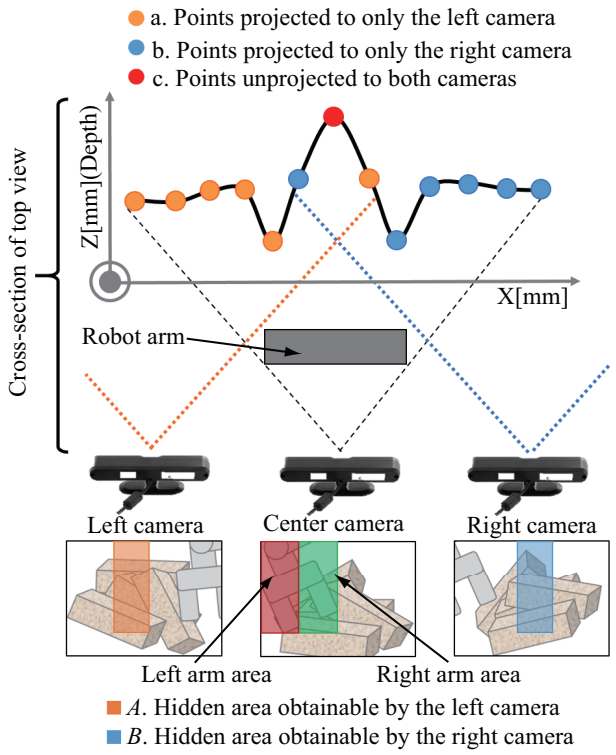


Fig. 4. Top view of experimental conditions. The points in this image indicate surficial shapes of the target work object occluded by the robot arm in the center camera and, they are classified into the specified three types.

camera was supposed to be far enough away from the background. Thus, the half-diminished reality image can only be composed by using homography transformation and blending the two images. However, when operating a robot, in must approach target work objects, and the background cannot be assumed two be a two-dimensional surface. Therefore, three-dimensional measurements are required to estimate the area occluded by the robot arm in the image.

III. Proposal Method

As mentioned above, our method requires three-dimensional environmental measurement for composition of half-diminished reality images. Thus, in this research, an RGB-D sensor is used for three-dimensional measurement. This sensor can provide both per-pixel RGB and depth information in an image. The depth information indicates the distance between an RGB-D sensor and a target work object.

A schematic view of the experimental conditions and positional relationships of the RGB-D sensors is shown in Fig. 2. In this research, three RGB-D sensors are installed, these three sensors are defined as the left camera, center camera, and right camera. Therefore, in a similar manner, the three images obtained by each RGB-D sensor are defined as the left camera image, center camera image, and right camera image, respectively. Three cameras are placed apart from each other to produce sufficient disparity among the images as shown in Fig. 2. In addition, a robot arm area in the center camera's image is defined as an arm area, and an area occluded

by the robot arm area is defined as a hidden area in each camera image (Fig. 3).

In Fig. 4, these points indicate surficial shapes of the target work objects occluded by the robot arm in the center camera, and they are classified into the following three types.

a. **Points projected to only the left camera**

b. **Points projected to only the right camera**

c. **Points unprojected to either cameras**

In general, conditional points are projected to the left camera image or the right camera image. However, if the target work objects have a rough concave and convex shape or the robot arm is approaching to the object, the point will not be projected to either camera as shown with the red point in Fig. 4. To reduce these points, many multiple view images are required. However, it is preferable to use the fewest cameras possible, considering the robot's size and the data traffic from each camera. The aim of this research is designing a system that use the minimum amount of cameras for three-dimensional measurement of the environment. Thus, three RGB-D sensors are sufficient to complete the proposed system. Although the resulting image loses some of its quality and details, the resolution is preserved and the composed image can be used to achieve our goal. As explained above, the following three processes must be preformed to compose the half-diminished reality images. occluded

A. **Acquisition of the arm area**

B. **Acquisition of the hidden area**

C. **Blending of the arm area and the hidden area**

In this paper, we present a method to detect the arm area from the center camera and estimate the hidden area from the others.

A. Acquisition of the arm area

The $\mathbf{w_{arm}} = [x_{\rm arm}, y_{\rm arm}, z_{\rm arm}, 1]^T$ is a three-dimensional arm position in the world coordinate system located around camera's center of the center camera as shown in Fig. 2. The arm position can be calculated using three-dimensional shape model data, and data from each joint angle of the robot arm. Thus, when the arm position projects the $\mathbf{m_{camC}} = [u_{\rm camC}, v_{\rm camC}, f_{\rm camC}, 1]^T$ in the image coordinate system, it can be calculated by the following equation.

$$\mathbf{m_{camC}} \simeq \mathbf{P}\mathbf{w_{arm}}, \tag{1}$$

where matrix P includes intrinsic parameters from the center camera, the relative position, and the orientation in the world coordinate system. This matrix can be obtained by camera calibration. Therefore, an arm area from the center camera can be calculated using the arm position in the world coordinate system in every frame.

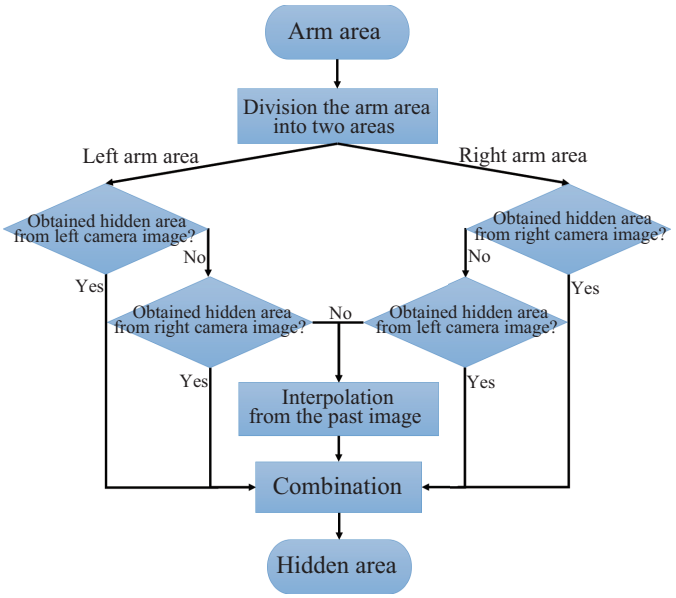


Fig. 5. Flow chart of procedures for obtaining the hidden area occluded by arm area in the center camera

B. Acquisition of the hidden area

The hidden area in the center camera is estimated from others camera images. First, the arm area is divided into two areas, and these two areas are defined as the left arm area and the right arm area. In this research, the hidden area is classified into three related types of surficial object shapes obtained by the depth information from left and right cameras (Fig. 4).

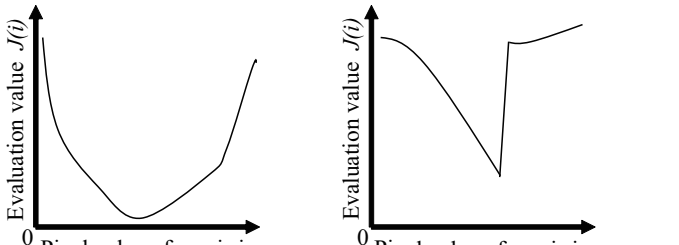
(A) **Hidden area obtainable by the left camera**

(B) **Hidden area obtainable by the right camera**

(C) **Hidden area unobtainable by either cameras**

Each method to estimate areas (A), (B), and (C) follows a procedure described below. In addition, the flow chart of our proposed method is shown in Fig. 5. First, the hidden area occluded by the left arm area is estimated from the left camera and the right arm area is estimated from the right camera to obtain the (A) and (B) areas, using the method in subsection Ⅲ-*A*. Second, pixels in the arm area that cannot be detected by the preceding process are reestimated by the camera on the opposite side. Finally, the (C) area, which cannot be detected by either camera is interpolated by using a previous image. The past image, in which the robot is not projected in the center camera's image, is captured by the center camera in advance. As a result we can compose half-diminished reality images.

1) Estimation of Hidden area: The $\mathbf{m}_{\rm cam} = [u_{\rm cam}, v_{\rm cam}, f_{\rm cam}, 1]$ is the three-dimensional background position in the world coordinate system located at the center camera's optical center as shown in Fig. 2. Our system is based on parallel-view stereo, which is achieved by locating each camera at a uniform height and optical axis direction. Thus, when the background position $\mathbf{w}_{\rm back} = [x_{\rm back}, y_{\rm back}, z_{\rm back}, 1]$ is occluded by the robot arm position $\mathbf{m}_{\rm cam} = [u_{\rm cam}, v_{\rm cam}, f_{\rm cam}, 1]$ in the center camera's image, if the background position $\mathbf{w}_{\rm back} = [x_{\rm back}, y_{\rm back}, z_{\rm back}, 1]$



(a) Case of non-dead-area (b) Case of dead-area

Fig. 6. Change the evaluation value. The left side graph shows the case of a non-dead area, and the right side graph shows the case of a dead area.

is projected to $\mathbf{m}_{\mathrm{camL}} = [u_{\mathrm{camL},i}, v_{\mathrm{camL}}, f_{\mathrm{camL}}, 1]$ or $\mathbf{m}_{\mathrm{camR}} = [u_{\mathrm{camR},i}, v_{\mathrm{camR}}, f_{\mathrm{camR}}, 1]$ in each image, the following equations are completed.

$$z_{\mathrm{back},i} - \frac{b_{\mathrm{camLC}} \cdot f}{(u_{\mathrm{camL},i} - u_{\mathrm{camC}})} = 0, \tag{2}$$

$$z_{\mathrm{back},i} - \frac{b_{\mathrm{camCR}} \cdot f}{(u_{\mathrm{camC}} - u_{\mathrm{camR},i})} = 0, \tag{3}$$

where f denotes the common focal length of all three cameras, which is a known quantity. Furthermore, the b_{camLC} denotes the distance between the left camera and center camera, and the b_{camCR} denotes distance between the center camera and right camera. The numerical subscript i in equations (2) and (3) denotes the i-th pixel of a horizontal line in each image. In this research, each pixel of all cameras contains RGB color and depth information, because RGB-D sensors are used. Thus, some hidden areas can be obtained by detecting the i values that minimize evaluation functions (4) and (5) in the horizontal line.

$$J_{LC}(i) = \left| z_{\mathrm{back},i} - \frac{b_{\mathrm{camLC}} \cdot f}{(u_{\mathrm{camL},i} - u_{\mathrm{camC}})} \right|, \tag{4}$$

$$J_{CR}(i) = \left| z_{\mathrm{back},i} - \frac{b_{\mathrm{camCR}} \cdot f}{(u_{\mathrm{camC}} - u_{\mathrm{camR},i})} \right|, \tag{5}$$

First, the left hidden area is estimated by the left camera, and the right hidden area is estimated by the right camera. When the left hidden area is estimated by the left camera, the preceding process is started at the left end point in the left arm area and, it calculates in the forward direction of the u axis as shown in Fig. 4. In the same manner, when the right hidden area is estimated by the right camera, the preceding process is started at the right end point in the right arm area and calculates in the reverse direction of the u axis. However, there are some pixels in each arm area that the preceding process cannot detect as shown in Fig. 4. In this paper, these pixels are defined as dead areas. Therefore, in this research, the following methods are used to detect whether the area is dead or not. For non dead areas, when some pixels in the left arm area are estimated by the left camera, the evaluation functions (4) and (5) take the extremely near zero value around taking the minimum value in evaluation value $J(i)$.However, for dead areas, these functions take high value as a whole and rapidly increase around taking the minimum value in the evaluation value $J(i)$. Thus, these features are used to detect whether the areas dead or not.

(1) The minimum value e of evaluation value $J(i)$

(2) The change rate e' around taking the minimum value of evaluation value $J(i)$

This change rate e' is defined in following equation.

$$e' = \max(J(d-1) - J(d), J(d+1) - J(d)), \tag{6}$$

where d denotes the minimum value determined by evaluation value $J(i)$, and e' denotes a pixel larger value between $J(d-1) - J(d)$ or $J(d+1) - J(d)$. When both the e and e' are lower than the threshold, it indicates that the pixel belongs a non-dead area. This threshold is defined heuristically. Conversely, when the e and e' are higher than the threshold, it means that the pixel belongs to a dead area.

Subsequently, if the non-dead areas are detected, these are reestimated by the camera on the opposite side. Finally, the areas that cannot be detected by the preceding process are interpolated by the following method.

2) Interpolation: In this research, we compose half-diminished reality images using images captured in real-time, which change from moment to moment. Thus, if the target work objects have rough concave or convex shapes or the robot arm is approaching them, it is still impossible to obtain some hidden areas that cannot be estimated by the preceding subsection's process. This is referred to as a deficient area. However, when operators control robots remotely, deficient areas cause decreased operating efficiency. Thus, in this research, the deficient areas that cannot be detected by the preceding subsection's process are interpolated by utilizing past images. Using this technique, we can obtain the hidden area.

C. Blending of the arm area and the hidden area

The arm area is blended with the hidden area in the center camera by composing the translucent arm area image. In this research, the following function is used to blend the arm area with the hidden area.

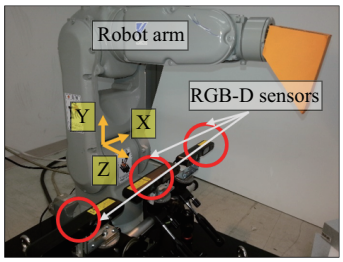
$$I_{output} = \alpha I_{arm} + (1 - \alpha) I_{cover}, \tag{7}$$

where I denotes RGB information of each pixel in each camera and α denotes the degree of transparency. $\alpha = 0$ indicates a perfect transparent image and $\alpha = 1$ indicates an opaque image.

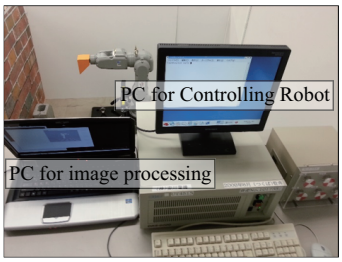
IV. Experimentation

A. Contents of Experimentation

We composed half-diminished reality images to test the proposed method with three images. In general, when operators control robots remotely in a disaster response environment,



(a) Positions about RGB-D sensors and the robot arm



(b) General view of this experimentation

Fig. 7. Images of the experimental environment. The left side image shows the positions of the RGB-D sensors and the robot arm. The right side image shows a general view of this experiment and PCs for image processing and controlling the robot.

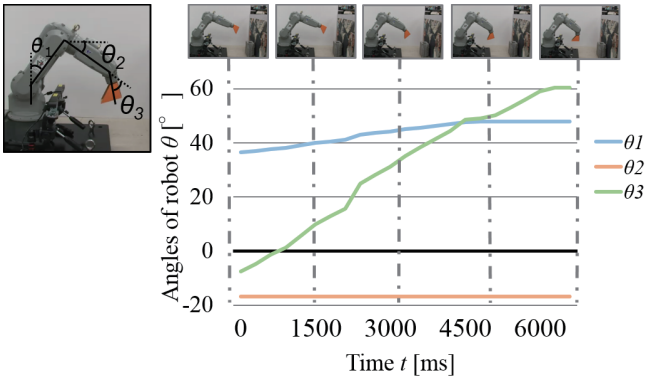


Fig. 8. This graph shows the orbit of the robot. Joint angles of the robot are defined in this image, and the robot copied debris the removal motion.

the robot works to remove debris. Thus, in this experiment, assuming that the robot removed debris, we composed half-diminished reality images in which debris occluded by the robot arm could be seen, thus allowing the operator to access the debris. Three RGB-D sensors（ASUS: Xtion Pro Live）and a robot arm（YASKAWA: MOTOMAN-HP3J）were used, as shown in Fig. 7(a). The hidden area occluded by the robot arm in the center camera was obtained from the left and right camera images, as described in Chapter Ⅲ.

Fig. 7(b) shows our experimental environment. Fig. 7(a) shows the world coordinate system and the relative positions of each camera and the robot. The center camera was located at the 150mm position in the Z axis and the 250mm position in the Y axis of the world coordinates. Furthermore, the left camera was located at the 200mm position and the right camera was located at the -200mm position in the X axis from the center camera position. The display resolution of all cameras was 640 x 480pixel. Fig. 9 shows the target work objects. In this experiment, to test debris removal by a remote controlled robot, we selected two object types to emulate actual debris. The first object type was a picture of a brick wall attached to a board; the board's size was 1000mm x 800mm. The second object type consisted of some cuboid blocks, similar to debris, scattered in front of the board; the size of the blocks was 200mm x 200mm x 100mm.

Three joint angles $\theta_1 \sim \theta_3$ of the robot arm were defined as

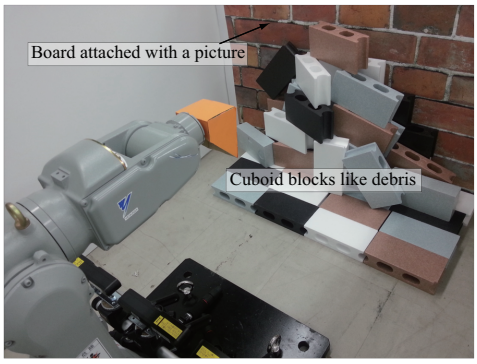
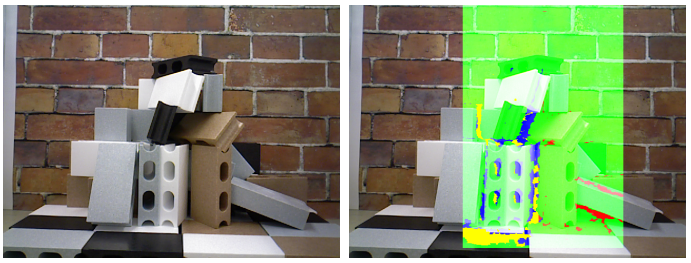


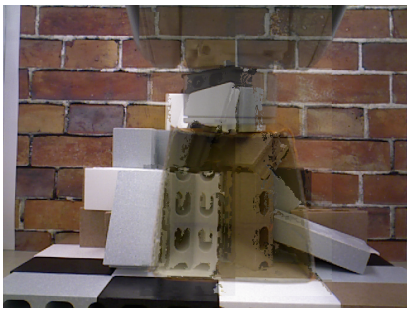
Fig. 9. Background objects. Two types of objects are used in the experimental environment. The first type was the board attached with a picture. The second type is comprised od cuboid blocks similar to debris scattered in front of the board.

(a) Left camera image (b) Center camera image (c) Right camera image

Fig. 10. Input images. Each camera image was captured when the robot arm moved following the orbit as shown in Fig. 8, at interval of $t = 4500$ms.



(a)Background image1 (b)Background image2



(c)Half-diminished reality image

Fig. 11. Output images. The upper left shows the past image which was captured in advance to interpolate the area which was not detected in the method of Ⅲ-*B-1*. Colors in the upper right image indicate hidden areas classified into three types in the process from Ⅲ-*B*. The bottom image shows the composed image.

shown in Fig. 8. It was assumed that the robot moved according to the orbit depicted in Fig. 8, and each camera image used when accessing the debris shown in Fig. 9 was an input image.

Furthermore, the numerical α was always 0.5 in all blending areas in equation (7).

B. Result

Each camera image captured when the robot arm moved, using intervals of $t = 4500\mathrm{ms}$, is shown in Fig. 10. The left camera image is shown in Fig. 10(a), the center camera image is shown in Fig. 10(b), and the right camera image is shown in Fig. 10(c). Furthermore, the past image, which was captured in advance to interpolate areas that were not detected by the method described in subsection Ⅲ-*B-1)* is shown in Fig. 11(a).

The output images from this experiment are shown in Fig. 11(b) and Fig. 11(c). Green areas in Fig. 11(b) denote the area that can be detected first by using the method from subsection Ⅲ-*B-1)*. Red and blue areas in Fig. 11(b) denote the areas that can be detected by the camera on the opposite side. Finally, the yellow areas in Fig. 11(b) denote the areas that were interpolated by the past image, and employing the method described in subsection Ⅲ-*B-2)*. By comparing Fig. 11(a) to Fig. 11(c), we confirmed that our proposed method can compose half-diminished reality images.

V. Conclusion

This paper presented a method to compose half-diminished reality images that can see through target work objects occluded by a robot arm, using three RGB-D sensors. This technique can be used to operate remote-control robots. In future work, we plan to evaluate the operating efficiency of using these half-diminished reality images.

References

[1] Satoshi Tadokoro: “Rescue Robotics Challenge”, Proceedings of the 2010 IEEE Workshop on Advanced Robotics and its Social Impacts, pp.92–98, 2010.

[2] S. Ali, A. Moosavian, H. Semsarilar, T. Kamegawa, and A. Kalantari: “Design and Manufacturing of a Mobile Rescue Robot”, Proceedings of the 2006 IEEE / RSJ International Conference on Intelligent Robots and Systems, pp. 3982–3987, 2006.

[3] C. Balaguer and M. Abderrahim: “Robotics and Automation in construction”, IEEE Robotics and Automation Magazine, vol. 9-1, pp. 1–20, 2002

[4] M. Moteki, K. Fujino, T. Ohtsuki and T. Hashimoto: “Research on Visual Point of Operator in Remote Control of Construction Machinery”, Proceedings of the 28th International Symposium on Automation and Robotics in Construction, pp. 532–537, 2010.

[5] N. Shiroma, N. Sato, Y. Chiu and F. Matsuno: “Study on Effective Camera Images for Mobile Robot Teleoperation”, Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication, pp. 107–112, 2004.

[6] S. Zokai, J. Esteve, Y. Genc, and N. Navab: “Multiview Paraperspective Projection Model for Diminished Reality”, Proceedings of the 2003 IEEE International Symposium on Mixed and Augmented Reality 2003, pp. 217–226, 2003.

[7] Y. Shen, F. Lu, X. Cao, H. Foroosh: “Video Completion for Perspective Camera Under Constrained Motion”, Proceedings of the The 18th International Conference on Pattern Recognition (ICPR2006), pp. 63–66, 2006.

[8] F. I. Cosco, C. Garre, F. Bruno, M. Muzzupappa and M. A. Otaduy: “Augmented Touch without Visual Obtrusion”, Proceedings of the IEEE International Symposium on Mixed and Augmented Reality 2009, pp. 99–102, 2009.

[9] B. Avery, C. Sandor and B. H. Thomas: “Improving Spatial Perception for Augmented Reality X-Ray Vision”, Proceedings of the IEEE Virtual Reality Conference 2009, pp. 79–82, 2009.

[10] P. Barnum, T. Sheikh, A. Datta, and T. Kanade: “Dynamic seethroughs: Synthesizing Hidden Views of Moving Objects”, Proceedings of the IEEE International Symposium on Mixed and Augmented Reality 2009, pp. 111–114, 2009.

[11] C. Sandor, A. Cunningham, A. Dey, and V.-V. Mattila: “An Augmented Reality X-Ray System Based on Visual Saliency”, Proceedings of the IEEE Virtual Reality Conference 2010, pp. 27–36, 2010.

[12] A. Enomoto and H. Saito: “Diminished Reality Using Multiple Handheld Cameras”, Proceedings of the 8th Asian Conference on Computer Vision 2007, pp. 130–150, 2007.

[13] T. Hashimoto, Y. Uematsu, H. Saito: “Generation of See-Through Baseball Movie from Multi-Camera Views”, Proceedings of IEEE International Workshop on Multimedia Signal Processing 2010, pp. 432–437, 2010.

[14] V. Buchmann, T. Nilsen, M. Billinghurst: “Interaction with Partially Transparent Hands and Objects”, Proceedings of the 6th Australasian User Interface Conference (AUIC2005), pp. 17–20, 2005.