# A Visual Sensor for Domestic Service Robots

Muhammad Attamimi and Takayuki Nagai

***Abstract*—In this study, we present a visual sensor for domestic service robots, which can capture both color information and three-dimensional information in real time, by calibrating a time of flight camera and two CCD cameras. The problem of occlusions is solved by the proposed occlusion detection algorithm. Since the proposed sensor uses two CCD cameras, missing color information of occluded pixels is compensated by one another. We conduct several evaluation to validate the proposed sensor, including investigation on object recognition task under occluded scenes using the visual sensor. The results revealed the effectiveness of proposed visual sensor.**

***Keywords*—Time of flight camera, visual sensor, camera calibration, occlusion detection, object recognition.**

## I. Introduction

In recent years, robots have been used in various scenes and applications. Under these circumstances, the appearance of robots coexisting with people is expected. In order to coexist with humans, the robot needs to grasp the actual environment in which humans live. In particular, it is important to recognize objects existing in the real environment. To realize that, an object recognition system that is robust to changes in the environment is required.

In realizing a robust object recognition system, the use of three-dimensional information is crucial because of its robustness to illuminating change. For that purpose, a sensor that can simultaneously acquire color information and the three-dimensional information is necessary. The development of three-dimensional measurement technology in recent years has been remarkable, and various methods have been proposed. These methods can be roughly divided into a method using a passive sensor and a method using an active sensor. In general, active sensors such as laser rangefinder (LRF) are less susceptible to illumination and the like, can measure distance faster, and more accurately than passive sensors such as stereo cameras. However, the three-dimensional LRF is large and very expensive, so it is not suitable for applications such as domestic service robots. Moreover, by driving a two-dimensional LRF by a motor, a method for obtaining highly accurate three-dimensional information is often used for environment mapping of a mobile robot. However, there is a problem that it takes time to scan in order to move the two-dimensional LRF with a motor. Time of flight (TOF) cameras are attracting attention as one of active sensors to solve these problems [1]. A TOF camera can acquire three-dimensional information at a speed of about 30 fps. Although its accuracy is inferior to LRF, it is very high compared to baseline stereo. The TOF camera is also a relatively small device, and it is easy to install in a domestic service robot.

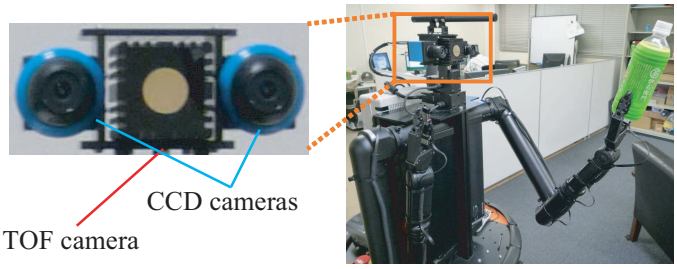


Fig. 1. A domestic service robot and the proposed visual sensor.

Therefore, in this paper, we construct a visual sensor for a domestic service robot that is able to measure information on the environment at high speed and with high accuracy; by integrating color information that is captured by the CCD camera and highly accurate three-dimensional information that is acquired from a TOF camera. In this case, by calibrating the TOF camera and the CCD camera, it is possible to match the corresponding pixel positions. However, since each camera has a different viewpoint position, occasionally, occlusion occurs when the point observed by the TOF camera is not observed by the CCD camera. At the pixel position where this occlusion occurs, a false correspondence between three-dimensional information and color information occurs, so distortion of an image such as a false contour occurs, which may adversely affect object recognition.

In this paper, we propose a method to detect occlusion at high speed and prevent erroneous response to this problem. Furthermore, with the proposed sensor, by using two CCD cameras, color information missing due to occlusion is compensated by another camera. This makes it possible to simultaneously obtain color information and highly accurate three-dimensional information in real time. Solving the occlusion problem in sensors is important for applications in object recognition. For example, a region hidden by occlusion includes information on colors and textures necessary for specifying the object. Experiments show that these problems can actually occur and that they can be solved by the proposed method.

TOF cameras have been drawing attention for several years and various applications are being studied. In [2], a TOF camera is used for three-dimensional map generation by the rescue robot. The estimation of 3D pose and camera motion using a TOF camera is proposed in [3], [4]. Moreover, the effectiveness of the TOF camera has been shown in [5], [6],

M. Attamimi is with the Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia e-mail: attamimi@ee.its.ac.id.

T. Nagai is with Departement of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, Tokyo, Japan..

Manuscript received April 19, 2005; revised January 11, 2007.

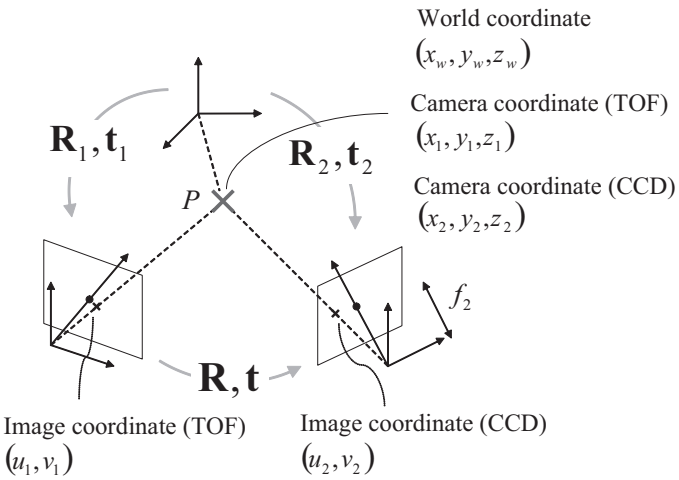


Fig. 2. Coordinate system transformation.

by applying the sensor to the tracking of the head and objects and the detection of people. However those studies focused only on three-dimensional information obtained by the TOF camera. The problem of calibration of a TOF camera and a CCD camera is described in [7], [8], [9]. However, the main focus of that study is the correction of the distance information of a TOF camera. Moreover, the problem of occlusion of the TOF camera and the CCD camera has been pointed out in [10], but it does not mention the concrete occlusion detection method, nor does it consider the problem in object recognition.

Recently, inexpensive three-dimensional sensors for games, Kinect [11], [12] are attracting attention. By using Kinect, it is possible to acquire color and depth information at the same time, but since the depth is measured by irradiating an infrared pattern, its accuracy is not as good as the TOF camera. Moreover, since it is supposed to be used for a game controller, there is a disadvantage that the distance near the sensor can not be measured. In addition, since only one CCD camera is used, it is impossible to compensate for missing information due to occlusion.

## II. Overview of proposed visual sensor

In this study, the proposed visual sensor is illustrated in Fig. 1. The sensor consists of a TOF camera and two CCD cameras. The visual sensor can acquire color and accurate three-dimensional information in real time by calibrating a TOF camera and two CCD cameras. To calibrate two different types of cameras, the following problems are required to be considered. First, the CCD camera has much higher resolution ($1024 \times 768$) than the TOF camera ($176 \times 144$). Second, each camera has its own parameters such as focal length, lens distortion, relative position, and so forth. The occlusion problem, which caused by relative positions of sensors, has to be resolved as well. These problems are discussed in the following sections.

### A. Time of flight camera

In general, the distance measurement capability of TOF camera is based on a TOF principle, i.e., the time taken for light to travel from an active illumination source to the objects in the field of view and return to the sensor is measured. In this paper, an off-the-shelf TOF camera SwissRanger SR4000 [13] is used. It emits a modulated near infrared (NIR) and the CMOS/CCD imaging sensor measures the phase delay of the returned modulated signal at each pixel. These measurements in the sensor results in a $176 \times 144$ pixel depth map.

### B. Camera calibration

In order to calibrate the visual sensor, we need to estimate the camera's parameters. In the geometric camera calibration, the parameters that express camera pose and properties can be classified into extrinsic parameters (i.e. rotation and translation) and intrinsic ones (i.e. focal length, coefficient of lens distortion, optical center and pixel size). The extrinsic parameters represent camera position and pose in three-dimensional space, while the intrinsic parameters are needed to project a three-dimensional scene onto the two-dimensional image plane.

We use Zhang's calibration method [14] in our proposed visual sensor, since the technique only requires the camera to observe a checkerboard pattern shown at a few different orientations. For the calibration of the TOF camera, the reflected signal amplitude can be used to observe the checkerboard

(a) (b) (c)

Fig. 3. An example of color mapping: (a) color image ($1024 \times 768$), (b) depth image ($176 \times 144$), and (c) mapped color image ($176 \times 144$).

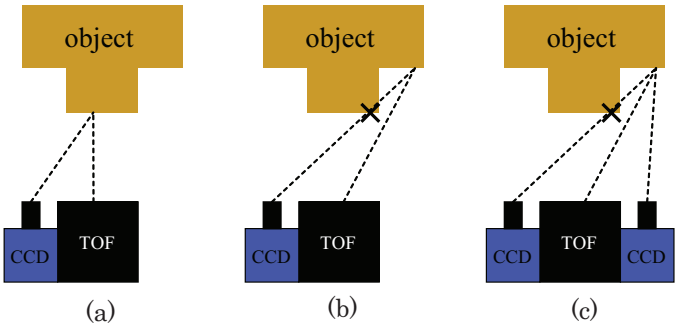


Fig. 4. An illustration of occlusion: (a) the point which can be measured by all sensors, (b) the point which cannot be measured by some sensor, and (c) compensation by the other sensor.

Fig. 5. Examples of occlusion.

pattern. Therefore, it is straightforward to apply the same calibration method.

### C. Color mapping

Assume that the geometric relationship between the camera coordinate systems is represented as in Fig. 2. Then, the coordinate of a point P can be written as,

$$\begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = [\mathbf{R}|\mathbf{t}] \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} \tag{1}$$

$$\mathbf{R} = \mathbf{R}_2\mathbf{R}_1^{-1}, \qquad \mathbf{t} = \mathbf{t}_2 - \mathbf{R}_2\mathbf{R}_1^{-1}\mathbf{t}_1 \tag{2}$$

where, $(\mathbf{R}_1, \mathbf{t}_1)$ and $(\mathbf{R}_2, \mathbf{t}_2)$ denote extrinsic parameters for TOF and CCD cameras, respectively. The rotation matrix and the translation vector for the CCD camera is represented respectively by $\mathbf{R}$ and $\mathbf{t}$.

It should be noted that lens distortions are assumed to be compensated by the estimated intrinsic parameters. Since the TOF camera provides $(x_1, y_1, z_1)$ directly, we can transform three-dimensional information for all pixels of the TOF camera into the CCD camera coordinate system using Equation (1). Finally, color mapping is carried out by the following perspective projection,

$$u_2 = f_2\frac{x_2}{z_2}, \qquad v_2 = f_2\frac{y_2}{z_2} \tag{3}$$

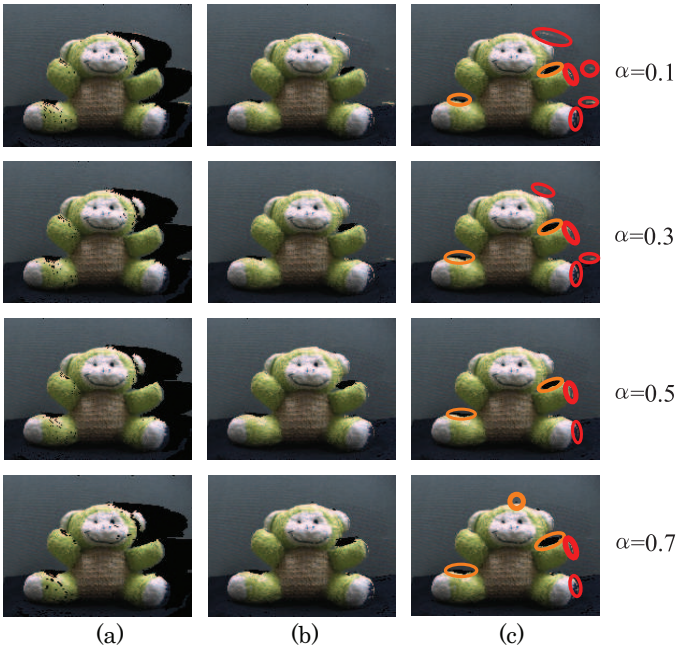


Fig. 6. Detection and compensation of occluded regions: (a) detection of occluded regions, (b) compensation of the occluded regions, and (c) explanation of compensated image: red and orange circles indicate remaining occluded regions and over-removed regions, respectively.

where $(u_2, v_2)$ and $f_2$ represent pixel location on the CCD's image plane and the focal length of the CCD camera, respectively. Figure 3 shows a result of the color mapping process.

## III. Occlusion

When integrating information of multiple sensors, it is necessary to consider a region in which some sensors can not measure it, i.e., an occlusion region. The following three regions are considered in this paper.

1) The region which is measurable from both of TOF and CCD cameras.
2) The region which is measurable only from the TOF camera.
3) The region which is measurable only from the CCD camera.

Since color information, which is captured by the CCD camera, is mapped to a corresponding pixels of the TOF's depth map, we have to take care of the second case.

Figure 4 illustrates the situation where occlusion occurs. Figure 4(b) depicts the situation where a point (marked with $\times$ in Fig. 4(b)) can be measured from the TOF camera but cannot be measured from the CCD camera. This situation leads to a false mapping of color information and a pseudo object appears as shown in Fig. 5. Therefore, occluded regions have to be detected and removed.

The Z-buffer method, which is widely used in the area of computer graphics, is applicable. However, it requires a three-dimensional data with polygon mesh representation. Instead,

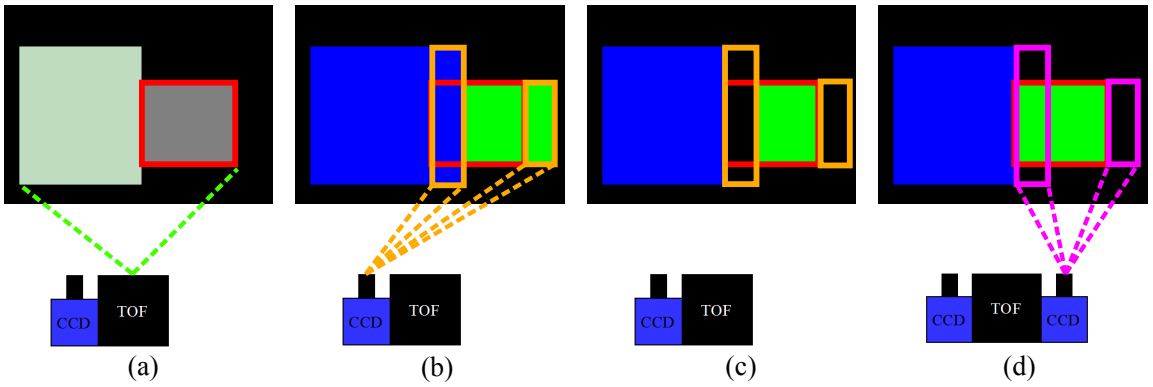


Fig. 7. An illustration of scene under occlusions: (a) a depth image with a target object (red rectangle), (b) the mapped color image with occluded regions (orange rectangle), (c) occluded regions are removed from the mapped color image, and (d) occluded regions are compensated (pink rectangle).

we propose a fast occlusion detection algorithm that works in the color mapping process. The basic idea is to use the property for alignment sequence of corresponding pixels between TOF camera and CCD cameras. If there is no occlusion, the corresponding pixel location of the CCD camera moves to the same direction as the horizontal scanning direction of the TOF camera's depth map. On the other hand, the corresponding pixel location of the CCD camera moves to the opposite direction when the occluded region starts. This is based on the property known as the order constraint [15] in stereo vision. In the three-dimensional restoration by stereo vision, the order constraint is used as a constraint in searching for corresponding points. In this study, however, the correspondence points are already known, and the order constraint is used for detecting the occlusion.

However, since the depth changes abruptly at the end position of occlusion (such as the boundary of the object), the accuracy of the depth information acquired by the TOF camera decreases. Therefore, if the determination of the end position of occlusion is used as it is, the occlusion region remains unremoved, and the phenomenon like a false contour occurs (see Fig. 6(c) red circle). This problem can be solved by expanding the regions to be deleted at the time of occlusion detection. The width to be expanded is set to a width proportional to the width of the occlusion region itself. This is because the width to be expanded is considered to depend on the amount of change in the depth. In other words, the larger the change in the depth, the wider the occlusion regions are needed. Since the amount of change in depth is considered to be reflected in the size of the occlusion region, the region is expanded by an amount proportional to the width of the occlusion region itself. The algorithm for detecting occluded regions is described as follows.

1) Find all pixels in the CCD camera that have correspondence with $N_u \times N_v$ pixels of the TOF camera; where, $N_u$ and $N_v$ denote respectively width and height of image captured by the TOF camera. Now, let $(U(u_i), V(v_j))$ be a coordinate of the CCD camera's image plane that corresponds to $(u_i, v_j)$ on the TOF camera. Set $i = 0$, $j = 0$, and $k = 0$.


Fig. 8. Objects used in the experiment of recognition under occlusions. Red circles indicate the differences in color and/or texture between each pairs.

2) Increase the value of $i$, and if $U(u_i) < U(u_{i-1})$ for $v_j - th$ row go to (3), otherwise go to (2), and if $i = N_u$ go to (5).
3) Increase the value of $k$, if $U(u_{i-1}) < U(u_{i-1+k})$ go to (4), otherwise go to (3). If $i + k = N_u$, then $u_i$ to $u_{i-1+k}$ becomes the occlusion region; and go to (5).
4) For $v_j - th$ row, $u_i$ to $u_{i-1+\lfloor k(1+\alpha)+0.5 \rfloor}$ is set as the occlusion region, and set $i = i - 1 + \lfloor k(1+\alpha) + 0.5 \rfloor$, $k = 0$, then go to (2). Here, $\alpha$ represents the rate at which the occlusion region is expanded, and $\alpha = [0, 1]$. Moreover, $\lfloor a \rfloor$ denotes the largest integer that does not exceed $a$.
5) Set $i = 0$ and increase the value of $j$, then go to (2). If $j = N_v$ the process ends.

It should be noted that the algorithm above is for the left-CCD and TOF cameras. For the right-CCD camera, the scanning has to carry out from right to left direction (from $i = N_u$ in decreasing direction). Figure 6(a) shows the result of proposed occlusion detection. It can be seen from Fig. 6 that, the results of detection are varied when the value of $\alpha$ is

set into 0.1, 0.3, 0.5, and 0.7.

Since our proposed system has two CCD cameras, which are mounted on both sides of the TOF camera, color information of occluded pixels are compensated by the other camera as shown in Fig. 4(c). Figure 6(b) depicts the occlusion compensation result. When the value of $\alpha$ is small, many occlusion regions are left out (red circles in Fig. 6(c)). On the other hand, increasing the value of $\alpha$ reduces the omission remaining in the occlusion regions. However, if the value of $\alpha$ is too large, the normal regions will also be removed (orange circle in Fig. 6(c)). Therefore, we should determine the value of $\alpha$ to balance the omission of the occlusion region with the elimination of the normal region. In this paper, we determine the value of $\alpha$ experimentally (see section IV-A).

## IV. Evaluation of Proposed Visual Sensor

### A. Accuracy of occlusion detection

In order to measure the accuracy of the proposed occlusion detection, we conducted an experiment. In this experiment, with respect to the image with occlusion (acquired from the left CCD camera), we manually labeled the occlusion pixel as the groundtruth. Then, we changed the value of $\alpha$ and calculated recall, precision, and F-measure of the proposed algorithm. When changing the value of $\alpha$ from 0 to 1, the best result was obtained with $\alpha = 0.5$. The value of recall, precision, and F-measure were $0.85$, $0.72$, and $0.72$, respectively. Hence, in this paper, we set $\alpha = 0.5$.

### B. The influence of occlusion on object recognition

One of our motivation in this study is to realize a sensor that can facilitate object recognition under occluded scenes. To validate the proposed sensor, we performed object detection in the occluded scenes and investigated the influence of occlusion on such a task. The definition of occlusion here is one occured in the region (2) as mentioned in section III (see Fig. 4(b)). The situation of the scene where occlusion occurs is illustrated in Fig. 7. First, we placed the object (red square of Fig. 7(a)) behind a certain object as shown in Fig. 7(a). When the occlusion regions are not considered, color information of some parts of the object (the orange squares of Fig. 7) is mapped falsely as illustrated in Fig. 7(b). On the other hand, if such regions can be detected and removed, it should be as shown in Fig. 7(c). Moreover, if we can compensate such regions by the other camera, it should be as illustrated in Fig. 7(d). Under these circumstances (Fig. 7(b), (c), and (d)), we performed object recognition.

Experiments were carried out using 10 objects shown in Fig. 8. These objects consist of five pairs of objects that are exactly identical in shape and differ only in some textures or colors. In other words, when different parts of color and texture are hidden by occlusion, it is difficult to distinguish pairs from each other. Different objects with only a part of these are relatively common in package of goods.

In this study, we adopted the object learning process proposed in [16]. In the learning phase, the user teaches the objects to the robot at various angles, and the robot generates a database which contains feature vectors of 50 frames per object. In the recognition phase, 40 frames of images, where occlusion occurred as shown in Fig. 7 were captured for each object, were recognized. An example of an actual scene is shown in Fig. 9. We used the method proposed in [17] to perform object recognition. We compared the recognition results of the following three cases.

1) Occlusions are not detected.
2) Occlusions are detected without compensation is not performed by the other CCD camera.
3) Occlusions are detected and compensation is performed by the other CCD camera.

The results are shown in Fig. 10. First for case (1), i.e., when occlusion is not considered (see Fig. 9(b)), information other than the object is included in the occlusion regions. This

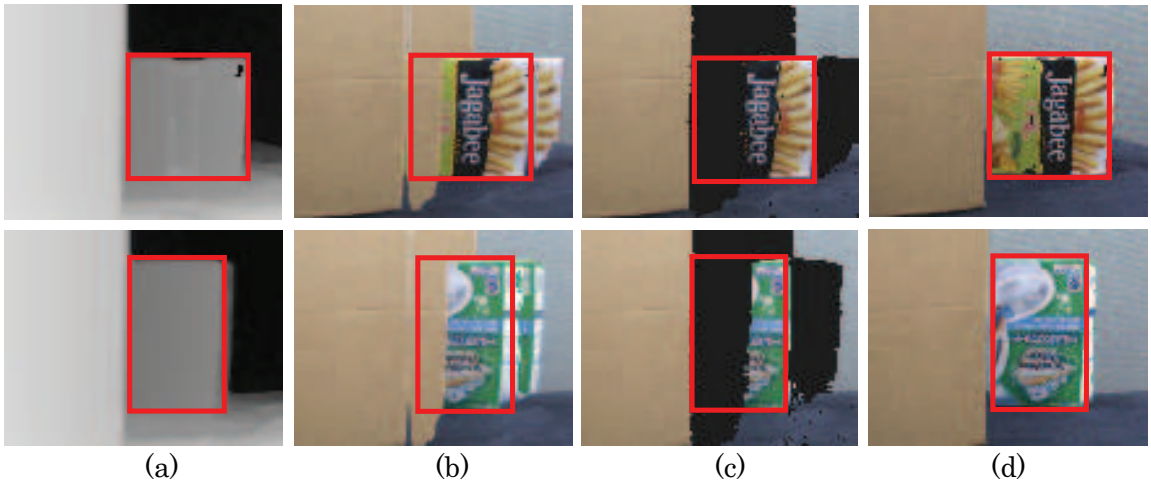


Fig. 9. Examples of scenes under occlusions: (a) depth images with target objects (red rectangle), (b) occluded regions are not detected, (c) occluded regions are detected, and (d) occluded regions are compensated.

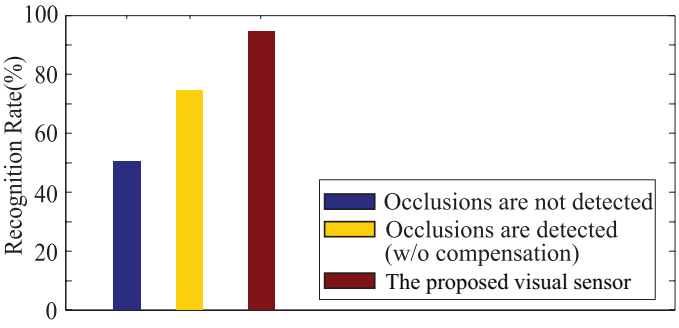


Fig. 10. Recognition results under occlusions.

fact leads to worse recognition rate (50%). Considering that there are two objects of the same shape, it can be said that this is the chance level of the recognition result. On the other hand, for case (2), i.e., when the occlusion region is detected and color information of the region is removed (see Fig. 9(c)), the recognition rate has improved to 75%. This is because the colors erroneously mapped in the object regions were removed, which enable to recognize the object when the minimum clues necessary for identification are remained. Furthermore, by using the proposed visual sensor (case (3)) (see Fig. 9(d)), occlusion by one camera is compensated by another camera. This compensation gained back the missing information due to occlusion, which can improve the recognition rate to 92.5%. This result is a special result using objects that are difficult to recognize due to the lack of some information. However, it is said that we are actually using objects that exist around us and that occlusion in the sensor has a possibility of adversely affecting object recognition. The proposed sensor is possible to avoid such adverse effects.

## V. Conclusion

In this paper, we have proposed a visual sensor which is based on the calibration of a time of flight camera and two CCD cameras. The proposed sensor is able to provide accurate three-dimensional information with registered color information in real time. We have evaluated the proposed visual sensor, and the results given in this study indicates that the proposed sensor can handle occlusion and contributed to the object recognition task. In addition, our visual sensor is also easy to implement on the domestic service robot.

## References

[1] T. Oggier, F. Lustenberger and N. Blanc, “Miniature 3D TOF Camera for Real-Time Imaging”, in Proc. of Perception and Interactive Technologies 2006, pp.212–216, June 2006.

[2] K. Ohno, T. Nomura and S. Tadokoro, “Real-Time Robot Trajectory Estimation and 3D Map Construction Using 3D Camera”, in Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.5279–5285, Oct. 2006.

[3] S. May, D. Droeschel, D. Holz, C. Wiesen and S. Fuchs, “3D Pose Estimation and Mapping with Time-Of-Flight Cameras”, in Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Oct. 2008.

[4] C. Beder, I. Schiller and R. Koch, “Real-Time Estimation of the Camera Path from a Sequence of Intrinsically Calibrated PMD Depth Images”, in Proc. of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol.XXXVII, pp.45–50, July 2008.

[5] S. B. Gokturk and C. Tomasi, “3D Head Tracking Based on Recognition and Interpolation Using a Time-Of-Flight Depth Sensor”, in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol.2, pp.211–217, July 2004.

[6] D. W. Hansen, M. S. Hansen, M. Kirschmeyer, R. Larsen and D. Silvestre, “Cluster Tracking with Time-Of-Flight Cameras”, in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp.1–6, June 2008.

[7] B. Bartczak, I. Schiller, C. Beder and R. Koch, “Integration of a Time-Of-Flight Camera into a Mixed Reality System for Handling Dynamic Scenes, Moving Viewpoints and Occlusions in Real-Time”, in Proc. of International Symposium on 3D Data Processing, Visualization and Transmission, June 2008.

[8] S. Fuchs and G. Hirzinger, “Extrinsic and Depth Calibration of TOF-Cameras”, in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–6, June 2008.

[9] Y. M. Kim, D. Chan, C. Theobalt and S. Thrun, “Design and Calibration of a Multi-view TOF Sensor Fusion System”, in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp.1–7, June 2008.

[10] A. Frick, B. Bartczack and R. Koch, “3D-TV LDV Content Generation with a Hybrid TOF-Multicamera Rig”, in Proc. of 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video, June 2010.

[11] Microsoft Kinect, http://www.xbox.com/en-us/kinect.

[12] PrimeSense, http://www.primesense.com/.

[13] MESA Imaging, http://www.mesa-imaging.ch/index.php

[14] Z. Zhang, “A Flexible New Technique for Camera Calibration”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.11, pp.1330–1334, Nov. 2000.

[15] H. H. Baker and T. 0. Binford, “Depth from Edge and Intensity based Stereo”, in Proc. of International Joint Conference on Artificial Intelligence, pp.631–638, Aug. 1981.

[16] M. Attamimi, A. Mizutani, T. Nakamura, K. Sugiura, T. Nagai, N. Iwahashi, H. Okada and T. Omori, “Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots”, in Proc. of IEEE Int. Conf. on Robotics and Automation, pp.745–750, May 2010.

[17] M. Attamimi, T. Araki, T. Nakamura, and T. Nagai, “Visual Recognition System for Cleaning Tasks by Humanoid Robots”, International Journal of Advanced Robotic Systems, Vol. 10, No. 11, pp. 1–14, 2013.