Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

Construction of an Intelligent Room Based on Gesture Recognition

Operation of Electric Appliances with Hand Gestures

Kota IRIE
Hitachi Ltd.
IBARAKI, JAPAN
kotairie@cm.jiji.hitachi.co.jp

Naohiro WAKAMURA
Chuo University
TOKYO, JAPAN
wakamura@sensor.mech.chuo-u.ac.jp

Kazunori UMEDA
Chuo University
TOKYO, JAPAN
umeda@mech.chuo-u.ac.jp

Abstract— **This paper proposes an intelligent room that is free of operator's position based on gesture recognition technologies. Intention and position of an operator are recognized by detecting hand waving, and pan-tilt cameras are zoomed and focused on the operator. The hand region is extracted using color information, and direction or number of fingers and motion of the hand region are detected. Home appliances such as a television set are controlled by using the direction or number of fingers and hand motions.**

Keywords: Intelligent Room, Gesture Recognition, Man-Machine Interface, Image Processing

I. INTRODUCTION

These days, computerization of our living environment is progressing with the development of IT technology. For example, networking of the home appliances is being realized and they are becoming more intelligent. On the other hand, the increase of their functions makes their operation complicated. Most of them are operated with buttons on them or a remote controller. With buttons, the position for operation is limited beside the appliance. With a remote controller, although this restriction is avoided, it is yet necessary to find it and go to pick it up since its position is not fixed. For such appliances frequently used in everyday life, intuitive operation is desirable for a user. Furthermore, it is also required that there is no restriction of the position to operate them. Therefore a non-contacting interface that uses man's natural actions is thought to be effective. Gestures, which we use frequently and intuitively in our everyday communication, are one of such man machine interfaces. Until now, many studies which recognize gestures from a sequence of images have been reported [1]-[3]. In some studies, gesture recognition technologies are used for an intelligent room, which is a room with functions of intelligent robots as recognition or interaction [4]-[6]. However, in many of the gesture recognition technologies, the place where gestures can be recognized is limited and thus they are not practical.

This paper proposes an intelligent room that is free of operator's position based on gesture recognition technologies. Home appliances such as a television set are controlled in the prototype.

II. OUTLINE OF THE INTELLIGENT ROOM

The intelligent room in this paper has some intelligent functions that are intended for a general office and a home. In the room, we can operate home appliances such as a television set and a lighting by gestures. The operator doesn't need a special attachment such as a glove or a microphone and can make operations in a natural state. The room specifies the operator autonomously even when two or more persons exist. The operator can make operations wherever in the room without any restriction. Gestures in the room are supposed to be by a hand or fingers.

The room carries CCD cameras, and detects an operator and recognizes his/her gestures autonomously with them. First, the system performs "detection of waving hands" with two cameras to detect the operator and then acquires three-dimensional (3D) position information. It carries out zooming of the cameras with the 3D information and restricts the region for detecting gesture recognition. Then it extracts a hand region using the color information. The skin color of the operator is registered in this stage, which improves the robustness of extracting the hand region to the difference of individual skin colors and the change of lighting environment. It performs "recognition of the number of fingers" using the features such as the area and the shape of the extracted hand region. Then it recognizes

0-7803-8463-6/04/$20.00 ©2004 IEEE

the gesture out of the registered ones. The recognized result is presented on a PC monitor and by a speaker, and interaction by the operator is made possible. Based on the recognized operation, a control signal is transmitted to the target appliance by the infrared remote controller connected to PC. Fig.1 shows the conceptual figure of the intelligent room.

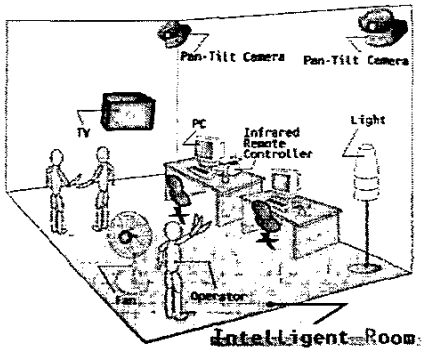


Figure 1. The conceptual figure of the intelligent room

III. DETECTION OF WAVING HANDS FOR DETECTION OF OPERATOR

In order to detect the person with an intention to operate an appliance, i.e. an operator, the intelligent room detects a waving hand. The method has already been presented in [4]. The outline of this method is shown below. The images are converted to low-resolution ones, and FFT is applied to each pixel of the low-resolution images. The detection of a waving hand's region is performed in them. The proposed method is robust to lighting condition and individual difference of skin color, because it doesn't use color information at all. Additionally, the method is very simple, because it doesn't require image processing for "recognition" of a hand.

The 3D position of the waving hand is measured by detecting it with two CCD cameras.

A. *Cyclic Change of Intensity Values of Images Caused by Hand Waving*

When a hand is waved, the intensity value of a pixel corresponding to the hand region vibrates between hand region and background. As a pre-processing, we make the resolution of the image lower. By this process, the pattern of the vibration is smoothed, and additionally, the robustness for noises is acquired and calculation cost is reduced.

B. *Application of FFT to Time Series of Intensity Values*

Each image is converted to low-resolution, and time scries of low-resolution images are obtained. Suppose the number of pixels of the images is $m \times n$, and $I(i,j,t)$ is the intensity value of (i,j) pixel (i=1,2,...,m,j=1,2,...,n) of t-th frame, as shown in Fig.2. Original image Fig.3(a) is converted to the low-resolution image Fig.3(b). The pixels in the rectangle of Fig.3(b) correspond to the region of the waving hand. The intensity value $I(i,j,t)$ of these pixels changes as illustrated in Fig.3(c), since the rate of the hand and the background changes periodically, according to the waving of the hand. As this change of intensity value is periodic with a constant cycle, we can utilize FFT for quantifying [7]. We apply FFT to intensity values of every pixels $I(i,j,t)$, and detect waving hands from the spectrum, illustrated as Fig.3(d). To remove the effect of noises like flicker of fluorescent light and reduce the calculation cost, FFT is applied to the pixels that satisfy

$$\max(I_{t-u+1},\ldots,I_t)-\min(I_{t-u+1},\ldots,I_t)\geq I_{dif}. \quad (1)$$

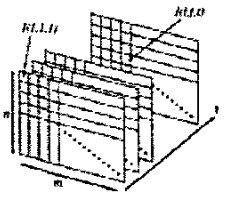
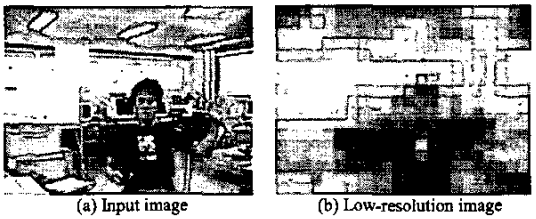
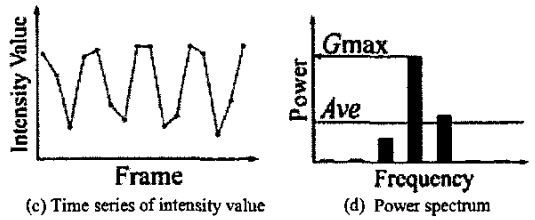


Figure 2. Time series of low-resolution images



(a) Input image (b) Low-resolution image



(c) Time series of intensity value (d) Power spectrum

Figure 3. Application of FFT to time series of intensity valies

C. *Recognition of Waving Hands by Discriminant Analysis*

Features are extracted from the power spectrum obtained from the time series of intensity values. We utilize the two features: the maximum value and the mean value of the power of the spectrum. For recognition of waving hands, the linear discriminant method[8] is applied to the feature.

D. Measurement of 3D Position by Stereo Vision

The 3D position of the waving hand is measured by detecting it with two CCD cameras.

1) The Calibration of the Cameras

The intrinsic and extrinsic parameters of the cameras are obtained by the following calibration method that uses a known pattern [9].

(1) Obtain the projection matrix between the 3D points and their projected points on a 2D image.

(2) Obtain the intrinsic and extrinsic parameter from the projection matrix. For the point on the image plane $\tilde{\mathbf{m}}=[u,v,1]^T$ and the point in 3D space $\tilde{\mathbf{M}}=[X,Y,Z,1]^T$, the projection equation for the perspective projection is expressed as

$$s\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{M}} = \mathbf{A}[\mathbf{R},\ \mathbf{t}]\tilde{\mathbf{M}} \tag{9}$$

where

$$\mathbf{A}=\begin{bmatrix} \alpha_u & -\alpha_u \cot\theta & u_0 \\ 0 & \alpha_v/\sin\theta & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{R}=\begin{bmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \mathbf{r}_3^T \end{bmatrix}, \mathbf{t}=\begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}. \tag{10}$$

When **P** is obtained, the matrix **A** that consists of intrinsic parameters, and the rotation matrix **R** and the translation vector **t** that are extrinsic parameters, are calculated.

2) Stereo Vision

3D coordinates of the matched point are obtained from the projection matrices of the two cameras. When the corresponding points $\mathbf{m}=[u,v]^T$ and $\mathbf{m}'=[u',v']^T$ are given, the following equations are satisfied.

$$s\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{M}},\ s'\tilde{\mathbf{m}}' = \mathbf{P}'\tilde{\mathbf{M}}. \tag{11}$$

From these equations,

$$\mathbf{BM} = \mathbf{b} \tag{12}$$

is introduced. 3D coordinates **M** is given by

$$\mathbf{M} = \mathbf{B}^{+}\mathbf{b}. \tag{13}$$

IV. SKIN COLOR REGISTRATION

In order to improve the robustness to the difference of individual skin colors and the change of lighting environment, we perform "skin color registration" as a pre-processing. For the color representation, Hue (H), Saturation (S) and Intensity (I) as illustrated in Fig.4 are used. In the HSI space, H and S are not affected by the change of brightness and suitable for the registration of the skin color. A cluster for the skin color region is formed in the HS space. At the recognition stage, the Mahalanobis distance from the cluster is calculated for each pixel and the pixels with the smaller distance than a threshold are recognized as a hand region. The average vector of the skin color cluster is set to $\mathbf{M}=[M_H,M_S]^T$, and a covariance matrix is set to **V**. The Mahalanobis distance d_M for a measured feature vector **X** is given by

$$d_M{}^2 = (\mathbf{X}-\mathbf{M})^T\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M}). \tag{14}$$

Fig.5 illustrates the Mahalanobis distance in the feature space.

Figure 4. HSI image

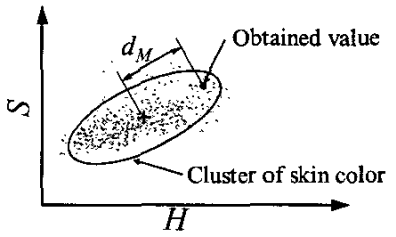


Figure 5. Distribution of HS value in feature space

V. GESTURE RECOGNITION

The hand region is extracted from a sequence of images by the method in Section IV, and a number of fingers, a pointing direction and operations by a hand are recognized for the sequence of the extracted hand region.

A. Recognition of the Number of Fingers

The number of the fingers is recognized by the following procedure, as illustrated in Fig.6.

For the extracted hand region based on the color information (1), erosion (2) and dilation (3) are applied. Then the region in (3) is subtracted from the region in (1), and the difference represents the finger regions (4). The number of the regions is recognized as the number of the fingers.

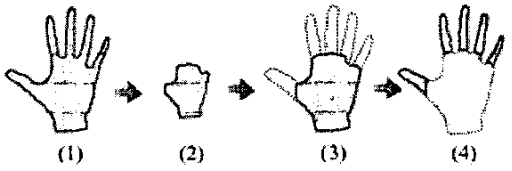


Figure 6. Recognition of number of fingers

B. Recognition of the Pointing Direction

Hand regions are extracted from the images captured by the two cameras based on the skin color information, and the principal axis in the image coordinate system is calculated for the extracted regions as shown in Fig.7.

The axis vector is expressed in the camera coordinate system as

$$\mathbf{a}_i = [1, \tan\theta_i, 0]^T . \quad (i = 1,2) \qquad (15)$$

The direction vector $\mathbf{A}_i$ from the origin of the camera coordinate system to the intercept of the principal axis $B_i(0,b_i)$ is given by

$$\mathbf{b}_i = [0, b_i, f_i]^T . \quad (i = 1,2) \qquad (16)$$

The normal vector $\mathbf{A}_i$ of the plane that is formed by the origin of the camera coordinate system and the principal axis is derived by the outer product of $\mathbf{a}_i$ and as,

$$\mathbf{A}_i = \mathbf{a}_i \times \mathbf{b}_i = [f_i \tan\theta_i, -f_i, b_i]^T , \quad (i = 1,2) \qquad (17)$$

where f_i is the focal length of each camera.

By multiplying $\mathbf{R}_i{}^T$, the transpose of the rotation matrix $\mathbf{R}_i$ of each camera in Eq.(10), each normal vector in the world coordinate system is obtained. Then the pointing direction by the finger is obtained as the outer product of the two normal vectors. That is, the pointing vector is obtained as the intersection of the two planes.

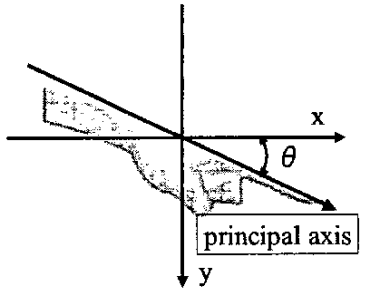


Figure 7. Recognition of finger direction

C. Motion Features of the Hand

In order to evaluate the hand motion, we use the sequence of motion vectors of the center of gravity of the hand region in the image plane. This feature is easily extracted with low calculation cost.

The motion vector $[dx_i, dy_i]^T$ is given by the following equations as illustrated in Fig.8.

$$dx_i = x_{i+1} - x_i , \qquad (18)$$

$$dy_i = y_{i+1} - y_i . \qquad (19)$$

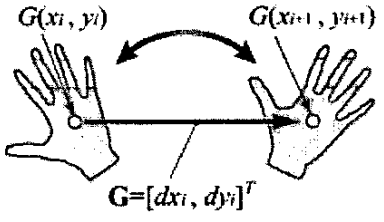


Figure 8. Movement vector of center of gravity

D. Gestures for Operating Appliances

Three gestures of "Cancel", "Up" and "Down" are introduced to operate the appliances. The gestures are recognized by using the number of fingers and the motion features as shown in Table 1.

TABLE I. GESTURES FOR OPERATION AND THEIR FEATURES

	Cancel	Up	Down
Number of finger	5	1	0
Feature	$\sum_k \lvert dx_k \rvert$	$\sum_k \lvert dy_k \rvert$	$\sum_k \lvert dy_k \rvert$

VI. OPERATION OF HOME APPLIANCES BY GESTURES

A. Indication of the Appliance for Operation

In order to operate an appliance, it is necessary to indicate it first. We use the "recognition of pointing direction" described in the previous section for this purpose. By pointing, a straight line is formed that has the obtained direction vector and passes the 3D position of the hand. The distance between the straight line and each candidate appliance is calculated, and the appliance with smallest distance is chosen as the appliance to operate. In the intelligent room constructed in this study, the home appliances for operation are a television set, a fan and a light, each of which can be operated by the infrared remote controller.

B. Operation of Home Appliances

After choosing the appliance, it is operated by gesture recognition. The operation method of each appliance is summarized in TABLE II.

TABLE II. OPERATIONS OF HOME APPLIANCES

Operation	Device	Gesture	Number of finger(s)
Power ON/OFF	TV, Fan, Light	Recognition of number of fingers	1
Channel	TV	Recognition of number of fingers n_i (i-th number)	Channel N=1~5 $\rightarrow N = n_1$
			Channel N=6~9 $\rightarrow N = n_1 + n_2$ $(n_1 = 5, n_2 = N - 5)$
			Channel N=10~12 $\rightarrow N = 10n_1 + n_2$ $(n_1 = 1, n_2 = N - 10n_1)$
Volume	TV, Fan	Recognition of the Up/Down gesture	1 or 0

VII. EXPERIMENTS

A. Experimental System

In this system, MVTec Halcon is used for image processing and other calculations and controls are performed by DELL PC (Pentium 4 2.2GHz). Images are acquired by two SONY EVI-D100 CCD cameras which carry a computer-controlled pan-tilt-zoom function. The two images from two cameras are composed by a Panasonic WJ-MS488 picture division unit and inputted into PC with a Leutron PicPort Color image capture board. In order to operate home appliances a Sugiyama Electron Crossam 2+USB infrared remote-controller is used that can learn commands and can be controlled by the PC.

B. Recognition of Waving Hands

We performed experiments of detecting waving hands by the method mentioned in Section III. Detection rates were evaluated for five subjects with different distances on the following conditions.

(1) Wave a hand at an arbitrary position in the camera's field of view for about 2[s].

(2) Suspend the hand waving for about 2[s], and then resume it.

They were repeated for 20 times, and when the hand waving was detected in the 2[s], we regarded the recognition of the hand waving was successful. TABLE III shows the recognition rate for different distances, and Fig.9 shows an example of the detection. It is shown that high recognition rate is realized for wide range of distance. The failure of detection typically occurred when the width of the waving hand was small.

TABLE III. RECOGNITION RATE

Distance	4m	5m	6m	7m	8m
Recognition rate	96%	96%	97%	92%	83%

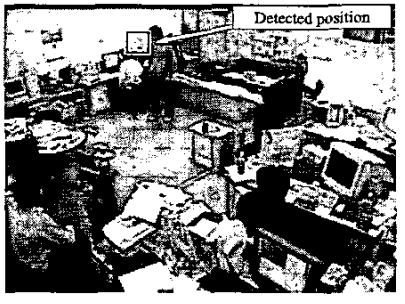


Figure 9. Example of the detection

C. Recognition of the Number of Fingers

We performed experiments of recognizing the number of fingers for five subjects by the method mentioned in Section V. First, skin color was registered for each subject as mentioned in Section IV. Then the subject changed the number of fingers from zero to five, and recognition rates for each number was evaluated. TABLE IV shows the recognition rate for different number of fingers, and Fig.10 shows the example of detection result. It is shown that high recognition rate is realized for each number of fingers. The failure of recognition occurred when, for example, the palm largely tilted upwards or downwards.

TABLE IV. RECOGNITION RATE

Number of fingers	0	1	2	3	4	5
Recognition rate	90%	98%	96%	82%	88%	88%

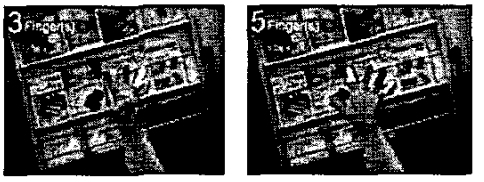


Figure 10. Experimental result

D. Recognition of the Pointing Direction

We performed experiments of pointing direction with the method mentioned in Section V. A subject stood on the position as shown in Fig.11. The angle for the two cameras was 90 [deg] on XY plane of the world coordinate system. Angle α between the camera 1 and the pointing direction was changed every 15 [deg] from 0 to 90 [deg], and angle β between Z axis and the pointing direction was fixed to 90 [deg], i.e. parallel to the floor. Fig.12 shows an example of detected pointing direction in the two images. Fig.13 and TABLE V show the results of the accuracy of detected pointing direction. It is shown that the pointing direction was roughly acquired, however the accuracy is low when the angle between the pointing direction and one of the two cameras are small. The reason is that with the small angle, the observed shape of the hand becomes nearly round and the accuracy of detecting principal axis becomes low.

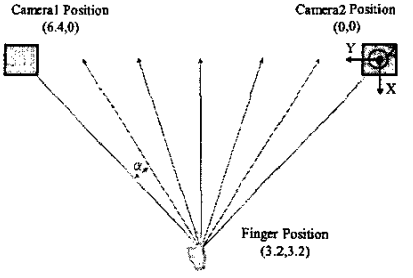


Figure 11. Experimental condition for pointing direction

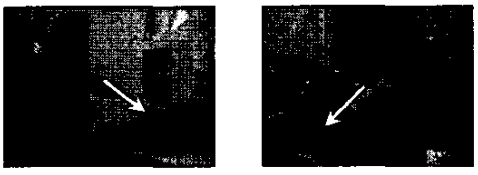


Figure 12. Example of detected pointing direction

TABLE V. AVERAGE AND STANDARD DEVIATION OF EXPERIMENTAL RESULTS

α[deg]	0	15	30	45	60	75	90
Average [deg]	17.3	25.1	33.3	43.4	59.2	77.4	90.0
Standard Deviation [deg]	5.0	1.1	3.5	1.0	2.7	3.0	10.6
β[deg]	90						
Average [deg]	95.9	93.8	90.2	91.4	95.0	89.2	87.0
Standard Deviation [deg]	2.1	1.3	1.1	0.7	1.9	1.9	4.3

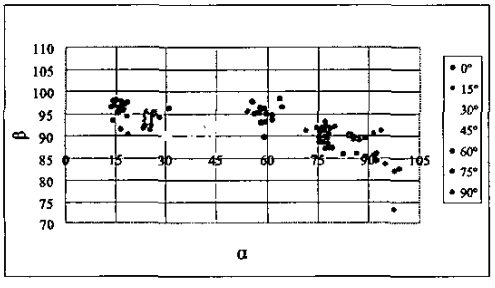
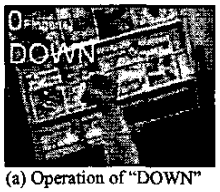


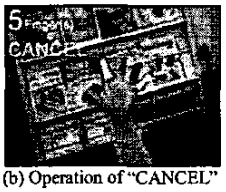
Figure 13. Experimental results

E. *Experiment of Operating Home Appliances*

We conducted the experiments to operate home appliances with the proposed methods. Fig.14 shows examples of recognized commands. Fig.15 shows the overview of the experiments. In the constructed intelligent room, we could operate the appliances correctly by only gestures with the proposed methods. The failure of the operation mainly occurred when the recognition of the number of fingers failed by the failure of skin color registration.



(a) Operation of "DOWN"



(b) Operation of "CANCEL"

Figure 14. Examples of recognized commands

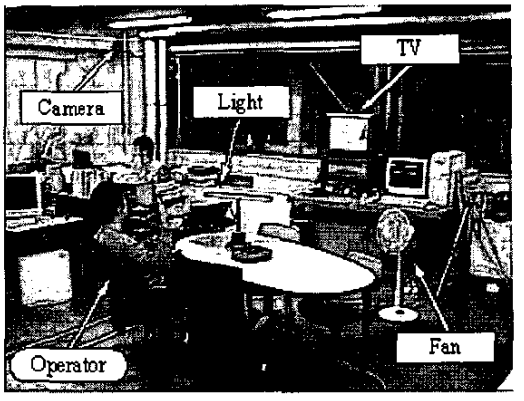


Figure 15. Overview of the experiments of operating home appliances

VIII. CONCLUSION

We proposed methods for recognizing gestures by a hand and fingers and built an intelligent room which uses them as the interface. Experiments showed the stability and robustness of the proposed methods. Future works include adding new gestures and improving the performance of the intelligent room. Introduction of an infrared camera or a mobile robot will expand the constructed system.

REFERENCES

[1] J. Sherrah and S. Gong, "VIGOUR: A System for Tracking and Recognition of Multiple People and their Activities", *Proc. of the International Conference on Pattern Recognition* (2000).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] P. Hong, M.Turk, T. S. Huang, "Gesture Modeling and Recognition Using Finite State Machines", *IEEE Int. Conf. on Automatic Face and Gesture Recognition* (2000).

[3] H. Wu, T. Shioyama, and H. Kobayashi, "Spotting Recognition of Head Gestures from Color Image Series", *Proc. of the International Conference on Pattern Recognition*, (1998) pp.83-85.

[4] Kota IRIE, Kazunori UMEDA: "Detection of Waving Hands from Images Using Time Series of Intensity Values", *The 3rd China-Japan Symposium on Mechatronics(CJSM)* (2002).

[5] Takashi Yamagishi, Yuichi Nakanishi, Kazunori Umeda:"Gesture Recognition for Realizing Intelligent Room", Mecatronics'01 5th Franco-Japanese Congress & 3rd European-Asian Congress Besancon(France), (2001), pp.293-298

[6] Taketoshi MORI and Tomomasa SATO, Robotic Room: "Its concept and Realization", *Robotics and Autonomous Systems*, Vol.28, No.2, (1999) pp.141-144.

[7] R. Cutler, L. S. Davis, "Robust Real-Time Periodic Motion Detection, Analysis, and Applications", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No. 8, (2000) pp. 781-796.

[8] Duda, R.O. and Hart, P.E. *Pattern Classification and Scene Analysis*, John Wiley & Sons (1973).

[9] Faugeras, O., Three-dimensional computer vision: a geometric viewpoint, MIT Press (1993).