Recognizing Human Behaviors with Vision Sensors in Network Robot Systems

Keiichi Kemmotsu, Tetsuya Tomonaka, Shigetoshi Shiotani, Yoshihiro Koketsu, and Masato Iehara
Advanced Technology Research Center, Mitsubishi Heavy Industries, Ltd.
1-8-1, Sachiura, Kanazawa-ku, Yokohama, Japan 236-8515
{keiichi_kenmotsu, tetsuya_tomonaka, shigetoshi_shiotani, yoshihiro_koketsu, masato_iehara}@mhi.co.jp

***Abstract* — A network robot system integrated with various types of robots via ubiquitous networks is a new concept which introduces an interactive robot living together with people. In this paper, we present a tangible network robot system composed of a mobile robot and vision sensors embedded in an environment, and show human behavior recognition methods necessary for providing diverse services desired by the people in good time. Typical human behaviors are described with logical sensors which are defined through data fusion processes spatially and temporally with the physical sensors of the mobile robot and those in the environment. Our system can be utilized to make human-robot communication and interaction friendlier and smarter.**

***Keywords*—behavior recognition, network robot, ubiquitous network, sensor fusion**

1. Introduction

Building a ubiquitous network infrastructure for our society by utilizing the latest information technologies (IT) is a key issue in realizing an invigorated, safe, secure, exciting, and convenient society in the 21st century.

On the other hand, interactive robots living together with people in non-industrial application areas have appeared recently. Their market size is expected to dramatically increase in ten years despite the fact that a conventional industrial robot market size has not grown in the past ten years [1]. Such new robot applications include various services and solution businesses, such as home use, health care, transportation, education, and so on.

Convergence of ubiquitous network technology and robot technology, both of which are the flagship technologies of Japan, would realize an innovative system, “network robots” [2]. The basic concept of the network robots is that various types of robots, which are called “visible robot”, “unconscious robot”, and “virtual robot”, are embedded in the ubiquitous network, and that diverse services would be realized through collaborations and interactions among those robots. The network robots would contribute to:

- creation of new life styles (wide spread and sophisticated services in life)
- solutions to address social problems, such as the aging of population and nursing care
- construction of a new IT society in the 21st century as a Japan-originated concept.

The network robots need to recognize the behaviors of persons whenever providing any service desired by the persons in good time. In this paper, we present a tangible network robot system composed of a visible robot (a mobile robot) and unconscious robots (environmentally embedded vision sensors) connected to each other via a network. The behaviors of a person are described through data fusion processes spatially and temporally with the vision sensors of the mobile robot and those of the environment. Recognized behaviors would break the ice in communication between the system and the person, and would help the system to decide what service the person would want.

The advantage of our system is that it is applicable in complex scenes where human behaviors could not be recognized by each individual sensor. Our intent is to apply the system to provide diverse services in a public space, such as a guide service, an information service, and a person search service in a shopping mall.

2. Typical Human Behaviors in a Public Space

In this section, we describe the typical behaviors of a person to be recognized when our system provides guide services in a public space as shown in Figure 1. We assume six key situations, and define human behaviors that the person would do under each situation as shown in Table 1. Our system should recognize those human behaviors and do some actions as shown in the rightmost column of Table 1. For example, the system recognizes a human behavior such as “waving one’s hand over the head toward a robot”, and then start a desired service.

Motion recognition methods have been developed for recognizing human gestures which are relatively short-term motions in previous researches [3], [8]. The methods often adopt a probabilistic state machine, such as hidden Markov models (HMMs), and they work well because the time-series patterns of those human gestures are moderately fixed.

On the other hand, our target behaviors in Table 1 include longer-term motions as well as short-term fixed motions, and thus the previous techniques are not sufficient for achieving our goal. The sequence and the frequency of short-term

motions should be observed and analyzed in order to recognize the longer-term motions. Our recognition method will be described in detail in Section 4.

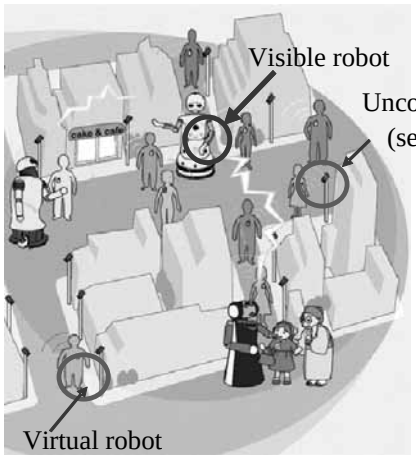

Figure 1. Guide service with a network robot system.

Table 1. A Person's Situations, Human Behaviors, and Robot Actions

A person's situations	The person's behaviors	Robot actions
Asking assistance	• Waving one's hand over the head toward a robot • Beckoning a robot • Talking to a robot	Starting a desired service for the person
Being in difficulty	• Looking around restlessly (Losing one's way) • Wandering (Looking for a person) • Having problems in using a machine	
Desiring information	• Looking at an exhibition or an advertising display	
Being busy	• Walking very fast • Passing through a crowd	No service
Expressing a will (Yes/No)	• Nodding one's head • Shaking one's head/hand	Continuing or stopping the service
Being in emergency	• Falling down in one's road • Lying on a road	Calling emergency

3. Configuration of Our Network Robot System

We propose a hierarchical structure for a network robot system consisting of three layers, that is, a physical sensor layer, a logical sensor layer and a human behavior recognition layer as shown in Figure 2. In the physical sensor layer, we use several sensors as shown in Table 2. The sensors are built in a visible robot, embedded in an environment, and attached to a human.

A single physical sensor cannot cover an entire space where the network robot system provides services for persons and also its accuracy and reliability vary with environmental conditions. For example, in the case of human detection and tracking with a single camera, we often encounter occlusion problems in image processing and fail to human detection and tracking. In the case of human detection with a radio frequency identification (RFID) tag, the detection range is limited within its transmitting range (typically several meters).

Logical sensors in the logical sensor layer denote virtual sensing devices that retrieve particular meaningful information from physical sensor data obtained via a network through data fusion processes spatially and temporally.

The human behavior recognition layer provides human behavior information necessary for controlling a robot action, for example, triggering a service, through fusion processes spatially, temporally and also semantically from logical sensor data. Because the logical sensors provide reliable and consolidated information with an adequate sensor placement, the human behavior recognition layer function can cover the entire service space.

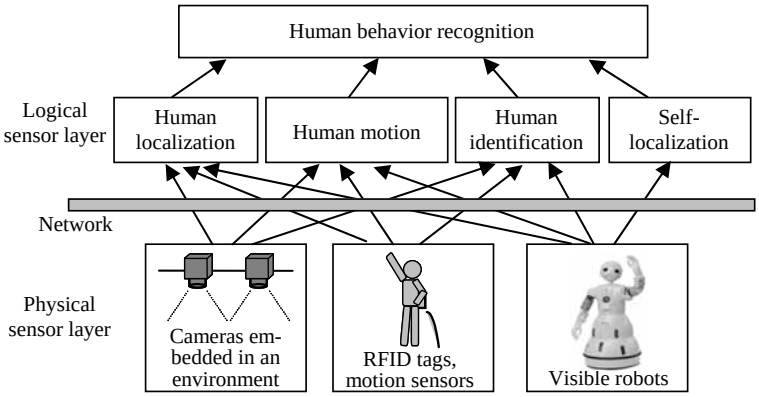

Figure 2. Three-layered structure of our network robot system.

Table 2. Examples of Physical Sensors

Type	Sensors
Built in a visible robot	Cameras, range sensors, touch sensors
Embedded in an environment	Ceiling cameras, floor sensors
Attached to a human (wearable sensors)	RFID tags, acceleration sensors

4. Examples of Logical Sensors

This section describes some examples of logical sensors: self-localization, human identification, human localization and human motion recognition. These logical sensors have been configured with cameras that are implemented on a visible robot and ceiling cameras. Figure 3 shows the overview of the visible robot, "wakamaru", developed by Mitsubishi Heavy Industries, Ltd., and its physical sensors. We use an omni-directional camera, a front camera, and odometry for self-localization, human identification, and human motion recognition. Odometry is the calculation of

robot position by the measurement of wheel rotation. Human localization is implemented with the ceiling cameras.

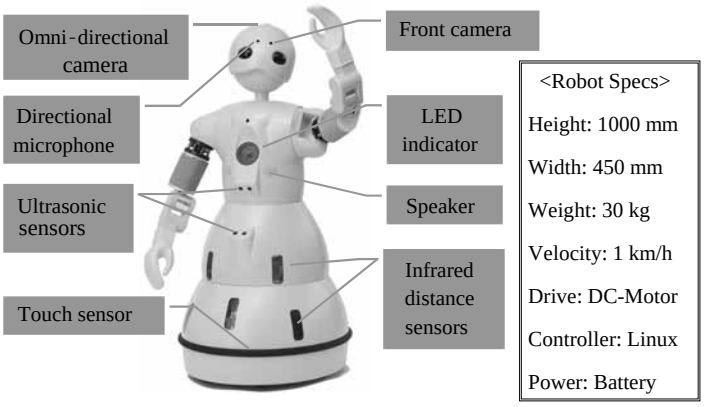

Figure 3. Sensors in a visible robot "wakamaru".

A. Self-localization

Self-localization is achieved by combining odometry with image measurement. An image including some landmarks attached on environments, for example, on the wall in a room is taken from an omni-directional camera as shown in Figure 4. The 3D positions of the landmarks are given beforehand. Then, the 2D positions of the landmarks in the image are extracted. Finally, the position and orientation of the robot are calculated by a probabilistic processing method [4], [5]. The result of self-localization shows that the position error is within 100 mm in 8 m by 4 m area as shown in Figure 5. The accuracy is drastically improved by combining odometry with the image measurement.

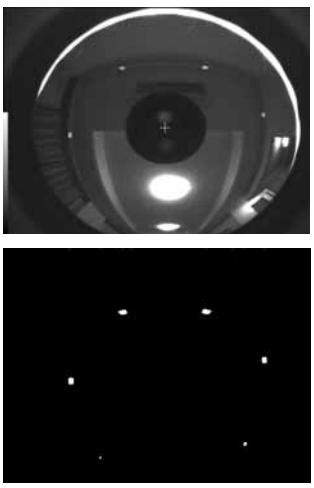

Figure 4. Omni-directional camera image (top) and detected landmarks (bottom).

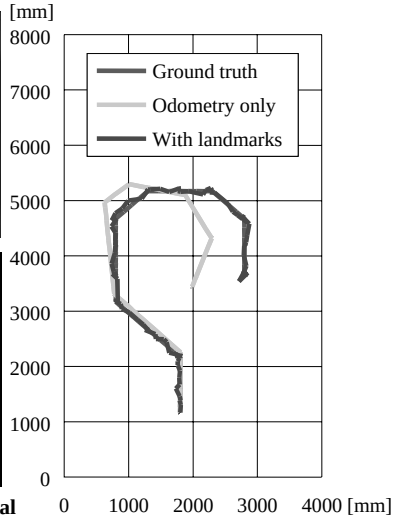

Figure 5. Self-localization result.

B. Human Identification

Our human identification process has three steps as shown in Figure 6. Firstly, a moving target (person) is detected from the image of an omni-directional camera. Then, the front camera is trained on the moving target and a face is detected from the image of the front camera with a face detection technique using shape and skin color information. Finally, a face recognition technique is used to identify the person.

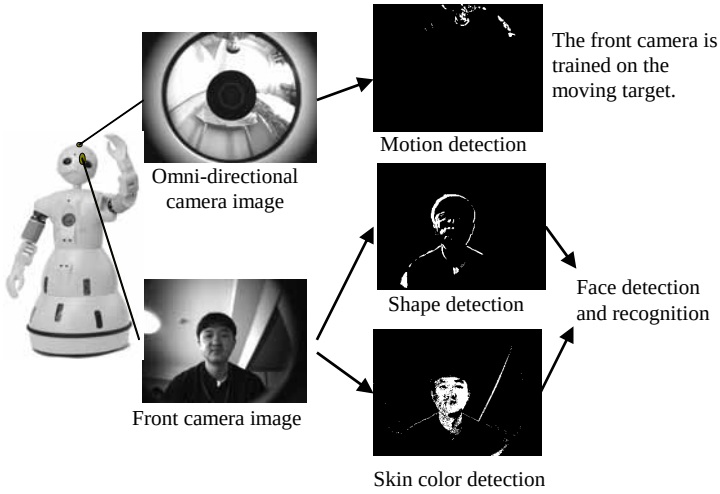

Figure 6. Face detection and recognition.

C. Human Localization

We use multi-viewpoint images captured by ceiling cameras to overcome a shadowing problem and an occlusion problem. A major advantage of adopting cameras for human localization is that they can track humans over a large area continuously.

In previous research [9], the N-ocular stereo image processing with multiple omni-directional cameras is used to detect the 2D positions of humans on a floor. In our system, the 3D position and the direction of humans are simultaneously estimated by a probabilistic modeling technique with a particle filter [4], [5]. This approach is not necessary to solve a stereo correspondence problem and it is easy to add more cameras to cover a larger area. Furthermore, our system estimates not only a human position but also basic human behaviors (walking, standing and sitting).

Figure 7 shows the sequence of human localization. Moving object regions are extracted from each image by a background subtraction method [10]. We use pixel color information to overcome illumination change. A background image is defined by a probability density function in a color space. An input image is converted into a probability image whose pixel intensity is a deviation from the estimated background image. Figure 8 shows the example of a calculated probability image. Brighter regions represent object regions which are moving with higher probability. Therefore, we track the brighter regions as humans using mean shift in an image sequence of each camera [11], [12].

Then, the human positions and their basic behaviors are estimated by integrating the tracked regions from the multiple camera images with a particle filter. Let $\boldsymbol{x}_t$ denote the state of

a human position and a basic behavior (walking, standing, or sitting) at time t. Let $\boldsymbol{z}_t$ denote the moving object region observation from camera images. $\boldsymbol{Z}_t$ is a history of $\boldsymbol{z}_t$, that is, $\boldsymbol{Z}_t = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t\}$. The goal of the particle filter is to estimate the posterior probability $p(\boldsymbol{x}_t|\boldsymbol{Z}_t)$ over the state variable $\boldsymbol{x}$ at time t.

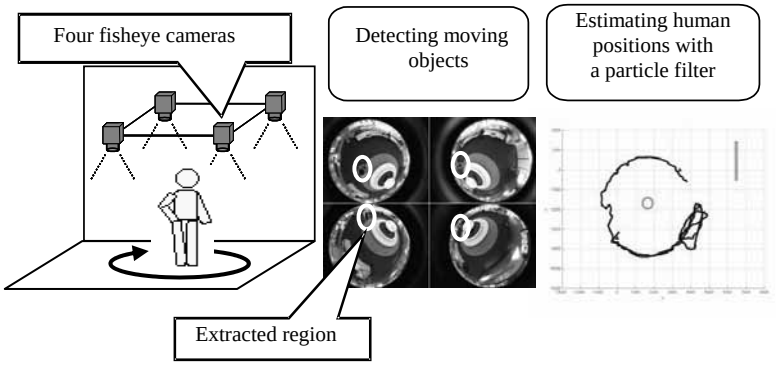

Figure 7. Human localization.

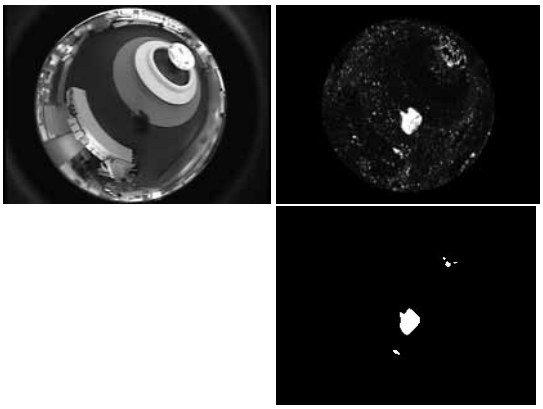

Figure 8. Moving object detection: an input image (top left), a probability image (top right), and extracted moving regions (bottom).

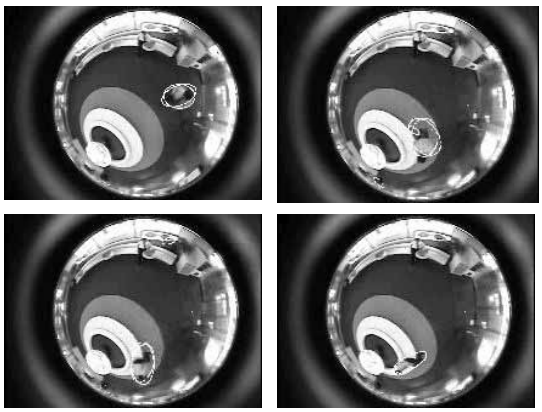

Figure 9. Tracking a human with one fisheye camera. A person is walking (top left and top right) and then sitting (bottom left and bottom right).

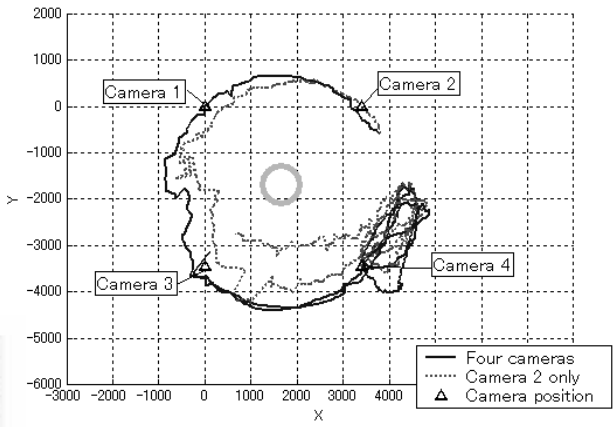

Figure 10. Experimental results (horizontal view).

Table 3. Basic Behavior Estimation Results

Basic behaviors	Estimation results		
	Walking	Standing	Sitting
Walking	75.5%	17.3%	7.2%
Standing	27.6%	70.0%	2.4%
Sitting	7.3%	4.1%	88.6%

In experiments, we took many scenes with four fisheye cameras. Figure 9 shows an example of tracking a person walking around. The accuracy of human localization is as shown in Figure 10. Multiple viewpoint image tracking is very stable compared with tracking with one camera. The basic behavior estimation success rate is more than 70% as shown in Table 3. There are very slow walking and swinging motions in the scenes and those cases sometimes result in wrong estimations.

D. Human Motion Recognition

Previous researches [3], [8] cover short-term gesture recognition, such as “lifting one’s right arm up”, “nodding one’s head”, and “waving one’s hand”. However, human behaviors we focus on include longer-term series, such as “eating”, “drinking”, “reading”, and “writing”. These motions last for at least several ten seconds, and often last for several minutes. Furthermore, any underlying short-term motion, such as “moving one’s right hand from around a table to one’s mouse”, is not a distinctive feature for recognition because such a short-term motion appears not only in “eating” but also in “drinking”.

Our idea is that we hierarchize a human motion into two motion classes according to the duration of the motion. One is a short term motion class and the other is a long term motion class. The short-term motion is defined as a motion that lasts for only several seconds, which has a moderately fixed time-series pattern. The long-term motion is defined as a motion that lasts for more than several ten seconds, which are composed of many short-term motions.

Figure 11. A sample scene where a person is drinking.

A human motion in a scene as shown in Figure 11 is tracked. Feature detectors (a face detector, a skin color detector and a blob tracker) detect the candidates of human body parts (the head, the right hand and the left hand) from an input image. The candidates of each body part are combined each other in a human model as shown in Figure 12. The human model is designed to represent a likelihood of "humanness" for each combination of the body part candidates. The combination given the highest likelihood is chosen as an estimated human body part position.

Short-term motions are recognized with the stochastic analysis of human body part motions using hidden Markov models (HMMs) as shown in Figure 13. The HMMs for several short-term motions are defined and trained. Short-term motions in an input image sequence are extracted with the trained HMMs. A long-term motion is recognized with a histogram analysis for a series of observed short-term motions as shown in Figure 14. The histogram of the short-term motions is compared with all the long-term motion histograms in a database, and the nearest one is chosen. Recognition results for some image sequences including four human behaviors are shown in Table 4. The recognition success rate is more than 80 %.

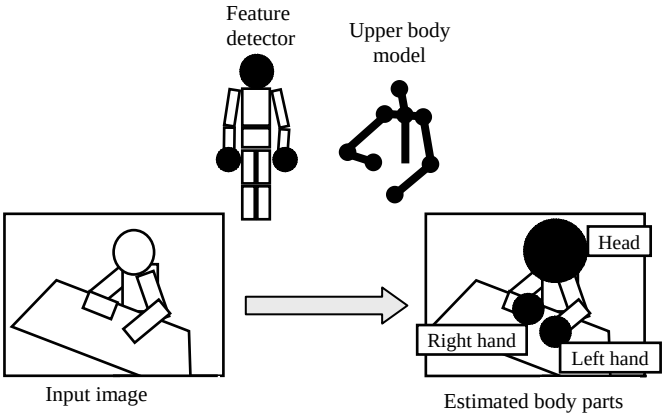

Figure 12. Estimating the positions of body parts.

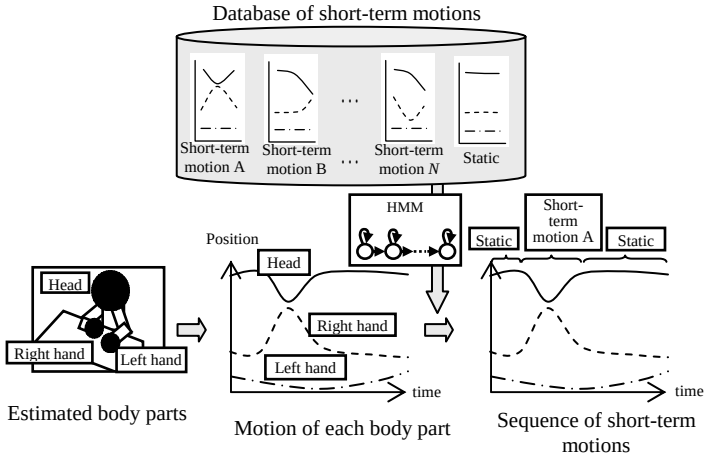

Figure 13. Recognizing a short-term motion.

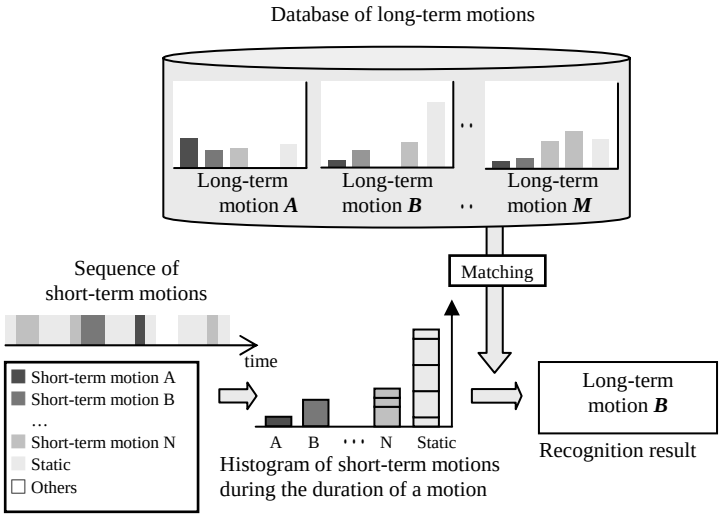

Figure 14. Recognizing a long-term motion.

Table 4. Human Behavior Recognition Results

Input motion	Recognition results			
	Eating	Drinking	Reading	Writing
Eating	100.0%	0.0%	0.0 %	0.0 %
Drinking	0.0%	100.0 %	0.0 %	0.0 %
Reading	2.1 %	6.3 %	91.7 %	0.0 %
Writing	0.0 %	2.1 %	14.6 %	83.3 %

5. Summary

The basic concept of network robots is that various types of robots are embedded in a ubiquitous network, and that smart services would be realized through collaborations and interactions among those robots. In this paper, we have presented a tangible network robot system composed of a

mobile robot and environmentally embedded vision sensors connected to each other via a network. And also, we have described how to recognize human behaviors with the network robot system. We have developed the following logical sensors:

- localizing a mobile robot
- identifying a human
- estimating human positions with multiple images
- recognizing human motions and behaviors.

We are now developing an advanced recognition system to classify 50 kinds of human behaviors with 90% accuracy at the Network Robot Project of Ministry of Internal Affairs and Communications in four years. We will use RFID tags and acceleration sensors in addition to the multiple fisheye cameras to recognize human behaviors in more complex scenes. The appearances of them are captured with the multi viewpoint images, and the positions and motions of the humans are acquired with RFID tags and acceleration sensors. These data are combined with a human model representing a human structure to identify the human behaviors. The system will be evaluated intensively for providing major robot services, for example, a guide service, an information service, and a person search service in a public space.

ACKNOWLEDGMENT

This research was supported in part by Ministry of Internal Affairs and Communications.

REFERENCES

[1] Japan Robot Assciation, Summary Report on Technology Strategy for Creating Robot Society in the 21st Century, 2001.

[2] Ministry of Internal Affairs and Communications, Final Report on Network Robot Technology, 2003 (in Japanese).

[3] T. Mori et al. “Human-like action recognition system using features extracted by human,” Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1214–1220, 2002.

[4] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, “Monte Carlo localization for mobile robots,” Proc. IEEE Int. Conf. on Robotics and Automation, 1999.

[5] J. E. Handschin, “Monte Carlo techniques for prediction and filtering of non-linear stochastic processes,” Automatica, vol. 6, pp. 555–563, 1970.

[6] M. A. Abidi and R. C. Gonzalez, Eds., Data Fusion in Robotics and Machine Intelligence, Academic Press, San Diego, 1992.

[7] M. Kam, X, Zhu, and P. Kalata, “Sensor fusion for mobile robot navigation,” Proc. IEEE, Vol. 85, No. 1, pp.108–119, 1997.

[8] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using Hidden Markov Model,” Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 1992.

[9] T. Sogo, H. Ishiguro, and M. Trivedi, “Real-time target localization and tracking by N-ocular stereo,” Proc. IEEE Workshop on Omni-directional Vision, pp.153–160, 2000.

[10] A. Elgammal, R. Duraiswami, D. Harwood and L. S. Davis, “Background and foreground modeling using nonparametric kernel density estimation for visual surveillance,” Proc. IEEE, Vol. 90, No. 7, pp.1151–1163, 2002

[11] D. Comaniciu, V. Ramesh, and P. Meer ,“Real-time tracking of non-rigid objects using mean shift,” Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2000

[12] R. Collins, “Mean-shift blob tracking through scale space, ” Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2003