

# Task Guided Attention Control and Visual Verification in Tea Serving by the Daily Assistive Humanoid HRP2JSK

Kei OKADA, Mitsuharu Kojima, Satoru Tokutsu, Yuto Mori, Toshiaki Maki, Masayuki Inaba  
Graduate School of Information Science and Technology, The University of Tokyo  
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan  
k-okada@jsk.t.u-tokyo.ac.jp

**Abstract**—This paper describes daily assistive task experiments that conducting on the HRP2JSK humanoid robot. We present overall action and recognition integrated system design to realize daily assistive behaviors autonomously and robustly, along with the demonstration that the HRP2JSK pours tea from a bottle to a cup and wash it after human drink it. To obtain autonomy and robustness, visual recognition and behavior control through perception information are important.

The significant issue tackled in this paper is what kind of task relevant knowledge is required for daily assistive task humanoids. Reducing search area is well-known technique to increase robustness, however, what kind of information should we embed in the robot is still the open problem, and in the humanoid case, the system has to cope with both manipulation and navigation task.

In this paper, we classified prediction based attention control based on following three task relevant knowledge: 1) Predicted search area to restrict potential object location in the recognition process, 2) predicted attention area to restrict image-processing area and 3) predicted visual features to eliminate mismatch. Task relevant knowledge is also used for vision guided behavior controls including 1) visual self-localization to recognize the position of a robot, 2) visual object localization to update the object location to generate behaviors and 3) visual behavior verification to confirm the success of the motion, are shown for adapting the planned motion to the real environment.

Finally, we demonstrated a tea service task by a humanoid robot. This task was repeated many times as presses or lab tourists demanded. Through this experience, we concluded that the robustness of the developed system reached to a satisfactory level.

## I. INTRODUCTION

Development of robotic behaviors in human daily environments is one of the most desperately-needed application [1]–[4]. Many researchers around the world address this problem with different approaches such as the behavior based approach [5], the tele-operation approach [6] and the cognitive learning approach [7] and so on. Among them, we have been developing a humanoid system based on knowledge based vision-guided robot system, which is archived through the development of three components: 1) Manipulation knowledge based whole body motion generation system [8], 2) Visual feature knowledge based 3D object recognition system [9], 3) Vision based environment and behavior verification system by using both manipulation and visual feature knowledge [10].

On the other hand, humanoid robots are expected to perform several application tasks at every occasion. Thus

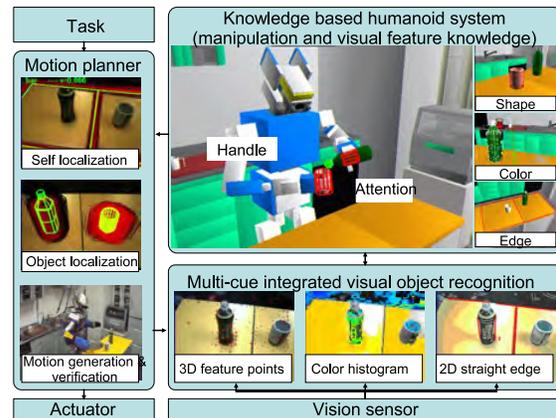


Fig. 1. Vision guided knowledge based humanoid robot system

a humanoid robot system is required to perform tasks with the high level of reliability.

The key technique to increase robustness is to guide visual attention and behavior control for reducing uncertainty and ambiguity. In this paper, we propose the use of task relevant knowledge for guiding visual attention and behavior control. In fact, we explored whether the manipulation and visual feature knowledge representation can be used for visual and behavior guiding. In this paper, we propose that the manipulation knowledge is used for guiding search area in the 3D space and the image plane. Visual feature knowledge is used for eliminating mismatch. This representation is also used for visual self localization, visual object localization and visual behavior verification.

We introduce a task knowledge based visual attention control method in the section III which navigates visual attention to the search attention area in the scene and the attention area in the view. In the section IV describes vision based behavior control including visual self localization, visual object localization and visual behavior verification. Section II describes the basis of our system and section VI presents tea serving task.

## II. ACTION AND RECOGNITION INTEGRATED HUMANOID SYSTEM

An overview of the knowledge based humanoid robot system is illustrated as the Fig.1. The system contains not

only geometric shape information of objects and environment but also contains manipulation and visual recognition knowledge.

#### A. Motion generation using manipulation knowledge [8]

We present how a humanoid motion planner works with manipulation knowledge. The sequence of an attention coordinates is the input of the planner. In the case of pouring tea behavior, the sequence represents a rotating motion of the top of the bottle using an attention coordinate of a bottle which is associated to the bottle as shown in the Fig.1. Then the planner calculates a motion of handle coordinates, which indicates the motion of the robot hand. Finally whole body motion is generated by calculating whole body joint angles from the motion of handle.

#### B. Multi-cue integrated object recognition using visual feature knowledge [9]

In order to recognize objects, we employ the Particle Filter algorithm [11], [12] which is widely used because of its robust characteristics. Each particle represents the hypothesis that indicates the 3D position of the target object and is weighted by likelihood using multi visual cue integration method [9]. The conditional density  $p(z_k|x_k)$  to calculate likelihood is represented as a following equation [13]:

$$p(z_t|x_t) = p_{point}(z_t|x_t) p_{color}(z_t|x_t) p_{edge}(z_t|x_t)$$

The position of the target object (state vector of the particle filter) can be written as  $\mathbf{x} = (x, y, z, roll, pitch, yaw)$  in a general manner.  $z_t$  is the measurement vector which indicates visual cues.

We have defined following three visual feature knowledge shown in the right top area of the Fig.1 includes Shape for calculating 3D distance between this shape and visual 3D feature points, Color for calculating similarity between this histogram and the histogram taken from the view images and Edge on an object surface for calculating 2D edge distance on the image plane. Please refer to the [9] for more detail.

#### C. Evaluation of multiple visual cue integration

Fig.2 shows that the multiple visual feature integration method provides robust object recognition. Top images (A) shows the result with 3D feature points. The red superimposed lines shows the recognition result and it shifts when occluded from the center to the left. The black superimposed lines presents particles. It can be seen that the particles are not converged.

Bottom images (B) shows the result by integrating 3D feature points and color histogram. Left bottom colored image shows the Hue images and Right bottom gray images indicates the likelihood of each pixel. By integrating color information, the system is able to track the target bottle while occlusion occurs.

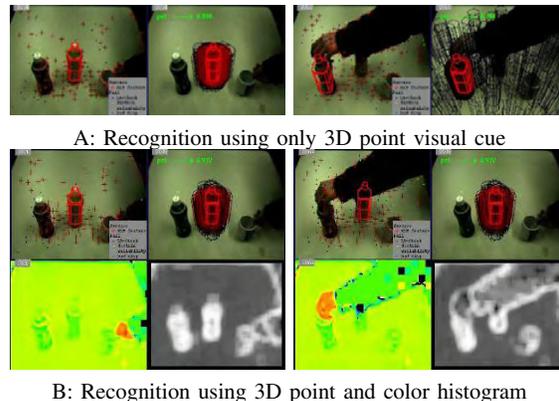


Fig. 2. Comparison of single visual cue object recognition and that of multi-cue integration

### III. TASK KNOWLEDGE GUIDED VISUAL ATTENTION CONTROL

For a daily assistive humanoid robot, it is important to control visual attention for realizing effective and robust object recognition. For example, when a robot find the cup, the system required to control visual attention in following three levels. 1) Directing gaze towards the potential location of the cup and search the target object on the location (Predicted Search Area). 2) Narrowing image view area to be processed where the cup features to be projected (Predicted Attention Area). 3) Predicting visual features to cope with occlusion using positional relationship between robot and the cup (Predicted Visual Features).

Since our robot system tightly integrates motion generation and visual recognition processes, the recognition process is able to predict the object location using task knowledge used in the motion generation process.

#### A. Search Area

We defined a 2D search area that particles are able to move along with the X and the Y axis (they are horizontal to the ground) and a 3D search area with the X and the Y axis and the yaw rotation (rotate around the Z axis). In the Fig.3, search area on the bar counter is the 2D search area and 3D search areas are located under the bar counter and the kitchen sink to recognize them. The red area below the kitchen tap also indicate the search area for recognizing rotational angle of the kitchen tap and water flow.

By introducing the search area, the robot is able to control it's gaze to the predicted target object position and the recognition process is able to limit the search space from the 6D (position and rotation in the 3D Space) to the 2D or 3D. This constraints enables us to realize practical object recognition system, since it is known that the particle filter with state space more than a few dimensions requires a large amount of particles that brings a slow convergence.

#### B. Attention Area

Narrowing area in the image view for image processing provides efficient and robust recognition. Fig.4 shows this

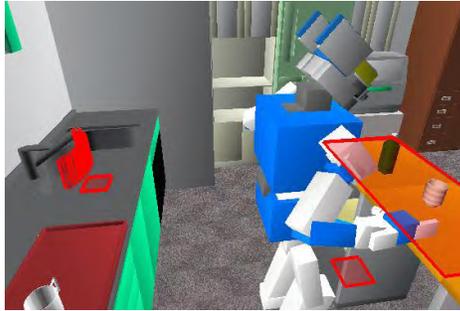


Fig. 3. Search area knowledge in the knowledge based system

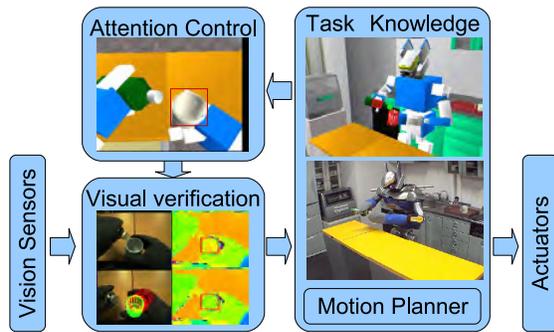


Fig. 4. Visual attention control in the knowledge based humanoid system

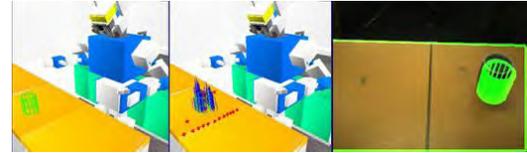
visual attention control mechanism using Attention coordinates in the knowledge described in the previous section. The image processing is applied to the attention area on the image plane where the this coordinates is projected. Instead of processing an entire image to detect the position of the cup and water in it, it uses restricted attention area for visual behavior verification such as searching the cup or find water flow using simple image processing method. See section IV-C for more detail in the image processing algorithms.

### C. Visual Features

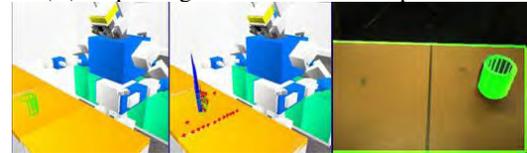
The above image(A) in the Fig.5 shows the object recognition result without using predicted visual features and the below image(B) presents the result using the prediction. The green colored cylinder object in the left column shows the 3D face model used for the object recognition. In the figure(A), all faces on the cylinder is drawn whereas occluded faces are not drawn in the figure (B).

Since the distance between faces of the model and visual 3D feature points are used for object recognition, occluded faces causes error. In the middle column, blue lines on the bar indicates the position of each particle and it's likelihood (weight). The particles has strong peak in the figure (B). In the figure(A) particles has multiple peaks. The right column shows the result of the recognition. Position errors about 1cm are observed in the figure (A).

We described in the case off 3D feature points here, this method is also used in the color histogram and 2D edge based object recognition.



(A) Cup recognition without view prediction



(B) Cup recognition with view prediction

Fig. 5. Comparison of accuracy in object recognition with respect to view prediction

## IV. TASK KNOWLEDGE GUIDED BEHAVIOR CONTROLS

In this section, we describe vision guided behavior controls for adapting the planned motion into the real environment.

Three visual behavior controls are required to perform each motion: 1) Visual self localization, 2) Visual object localization, 3) Visual behavior verification.

Before performing each motion, the robot is assumed to be located on the `spot` position and positions of task relevant objects are known in advance. Thus the visual self localization and the visual object localization are required.

After the motion, the robot verifies the behavior. We classified the verification process into two groups. One is an indirect verification and another is a direct verification.

### A. Vision based self localization based on visual feature knowledge

Fig.6 shows the vision based self localization method. In order to perform the tea serving task, the robot is assumed to be located on the `bar counter spot`. An associated object to this `spot` is yellow colored bar counter table, which is shown in the top left image. Edge knowledge is used for recognizing the bar counter. Fig.7 shows the case of self localization using sink object model with the Edge visual feature.

Visual recognition process calculates the relative coordinate between an actual robot position in the real environment and the `spot` position in the model environment. Then, it update the current robot position and the robot walks in order to maintain consistency with the model world. Since this walking action produces translation error, visual self localization process usually repeated few times until convergence errors lower than 1[cm].

### B. Vision based object localization based on visual feature knowledge

Visual object localization process updates the object position in the model environment along with the actual object position. In order to recognize the object, it uses the visual feature knowledge which is associated with the object model.

Fig.8 shows that a cup and a plastic bottle are recognized by using the proposed method. They have the visual feature

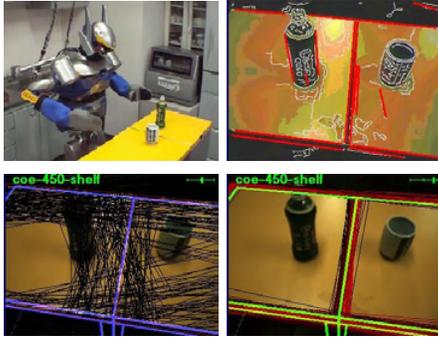


Fig. 6. Vision based self localization using counter knowledge.

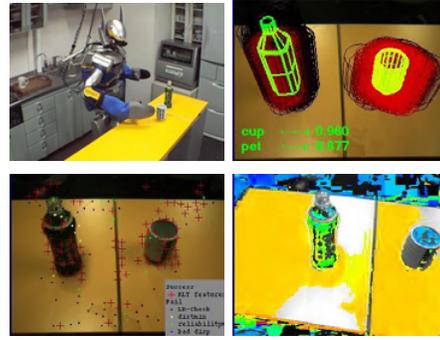


Fig. 8. Vision based cup and plastic bottle recognition

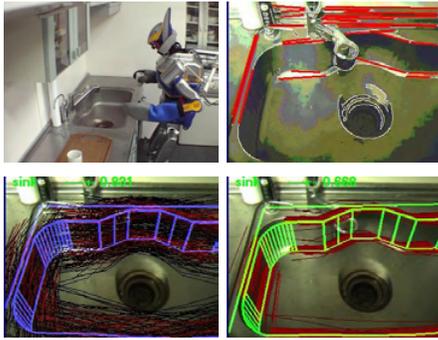


Fig. 7. Vision based self localization using sink knowledge.

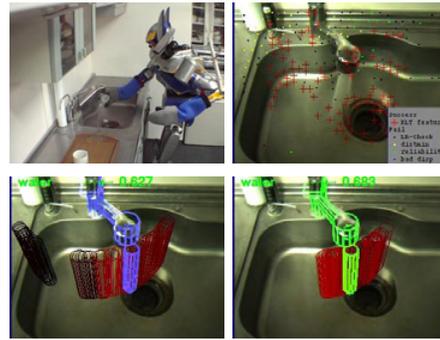


Fig. 9. Water flow recognition using tap knowledge and 3D features points

knowledge as described in the previous section. The cup has shape information and the plastic bottle has both shape and color histogram information. Bottom images in the Fig.8 shows 3D feature points and color hue image. Top right image shows recognized positions of objects. The holding cup motion is generated upon this information as shown in the top left image.

### C. Vision based behavior verification using task relevant knowledge and visual feature knowledge

After the motion execution phase, we use vision based behavior verification process to confirm the success of the motion. We classified the verification process into indirect verification based on task relevant knowledge and direct verification which uses object relevant knowledge.

1) *A direct verification with visual feature knowledge:* A direct verification examines the success of the behavior using knowledge associated with the target object. For example, in order to verify the cup holding behavior, recognition of the cup in the hand is required.

Fig.10 shows an example of the direct verification of the cup and the bottle holding behavior. It uses 2D directed edges(Edge), which is illustrated as red point in the left image, to calculate the cup position in the robot's hand. The bottom row images shows the grasping the plastic bottle behavior verification through the bottle recognition. It uses the Shape visual feature of the cap of the bottle.

2) *An indirect verification with task relevant:* An indirect verification examines the success of the behavior using

knowledge associated with the task. For example, in order to verify the pouring tea behavior, the robot examines if there exist tea in the cup or not.

In order to detect tea in the cup, we use color a histogram based recognition method. Images on the left column in the Fig.11 present Hue information. The middle and the right column correspond to the Saturation and the Intensity image. Images taken before the tea pouring behavior are shown in the top row and images after the behavior are listed in the middle row.

Graphs in the bottom row shows the change of histograms before and after the behavior. Red rectangles in the upper images present an area to calculate histograms, which is determined by projecting the cup position on the view image plane. These graphs indicates that the existence of tea drink in the cup is recognized using the change of the histogram. Similarity is calculated using distance between two color histogram by using the Bhattacharyya coefficient.

This method can be apply to any liquids with the color, how ever difficult to detect clear liquids as water.

Recognizing water shown in the Fig.9 is applied to verify the open and close tap behaviors. Water is modeled as a cylinder object coupling with the water outlet object. The position of the water model is constrained by the water outlet joint model and the recognition process calculates the similarity between water model and visual information using distance between 3D feature points and cylinder faces.

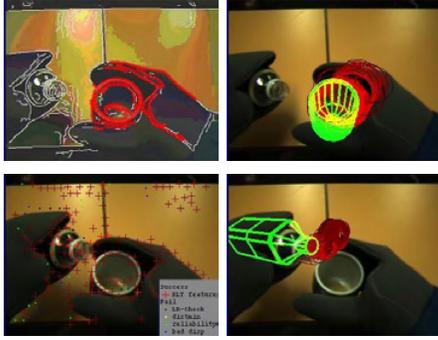


Fig. 10. Behavior verification using visual feature knowledge.

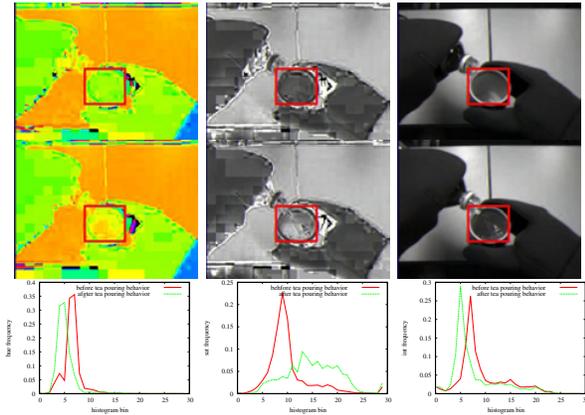


Fig. 11. HSI images and histogram changes in pouring tea behavior

## V. TASK-LEVEL PLANNER FOR SCENARIO DESCRIPTION

Since our system has capable of providing high level autonomous behaviors. It is easy to connect high level task planner for describing and controlling scenario of the robot task. See [14] for more detail.

We adopt the STRIP type operator for each behavior. For example the HOLD operator has preconditions (ON ?OBJECT ?SPOT) (AT ?SPOT), action (HOLD ?OBJECT) and effects (HOLD ?OBJECT) !(ON ?OBJECT ?SPOT), POUR-TEA operator as (HOLD CUP) (HOLD BOTTLE) (AT BAR) precondition, (POUR TEA) action and (POURED CUP) effects, and WASH-CUP operator has (HOLD CUP) (AT SINK), (WASH-CUP) and (WASHED CUP)

Thus, the first half of the demonstration scenario described in the next section can be generated by giving (POURED CUP) as the goal status to the planner, and the last half can be generated by (WASHED CUP) (ON CUP SINK).

## VI. TEA SERVING TASK

In this section, we describe the description required for demonstrating the tea service task. This humanoid task is a part of the demonstration to show the the accomplishment of “21st Century COE Information and Technology Strategic Core: The Real-world Information System Project” [15] and was repeated many times as presses or lab tourists demanded.

Behaviors	Visual controls	Object	Knowledge
Behaviors with self localization		Cup	Shape
Move to counter	Recog. counter	Bottle	Histogram, Shape
Move to kitchen	Recog. sink	Counter	Edge
Behaviors with object localization		Sink	Edge
Hold a cup	Recog. cup	Search area	Ttarget
Hold a bottle	Recog. bottle	On counter	Cup, Bottle
Place a cup	Recog. cup	Counter foot	Counter
Place a bottle	Recog. bottle	Sink foot	Sink
Behaviors with visual verification		Under tap	water flow
Pour tea	Recog. tea	Event	Knowledge
Open tap	Recog. water	Recog. tea	Color histogram
Close tap	Recog. water	Recog. water	Water flow model
Wash cup	—		

TABLE I

KNOWLEDGE DESCRIPTION IN THE KITCHEN EXPERIMENT.

### A. Task scenario

We demonstrated the tea serving task experiment as shows in the Fig.12. The scenario of this experiment is as followings. The number on each line corresponds to the number in the figure.

- 1) The robot recognizes the cup (1) and holds (2).
- 2) The robot recognizes the bottle (3) and holds (4).
- 3) The robot pours tea into the cup from the bottle (5).
- 4) The robot places the cup (6) and the plastic bottle (7).
- 5) The human drink tea in the cup and place it (8-9).
- 6) The robot recognizes the cup (10) and holds.
- 7) The robot walks to the kitchen (11-12).
- 8) The robot localizes self position (13).
- 9) The robot opens the tap (14) and conform it (15).
- 10) The robot washes the cup (16).
- 11) The robot closes the tap and place the cup.

### B. Knowledge description

This section describes knowledge required to perform the experiment. Behaviors required for archiving this experiment are following 10 units as listed in the left table in the TABLE I. First two behaviors require a vision based localization, Next four requires an object detection and the last four requires behavior verification.

We defined four object in the demo scene. For each object, we described associated visual recognition knowledge as shown in the right top table. Recognition of the cup and the plastic bottle, the bar counter and the kitchen sink are presented in Fig.8, Fig.6, Fig.7 respectively.

The right middle table indicates “Search Area” which we defined for this experiment as in the Fig.3. Three search areas are defined for detecting the objects for grasping(cup and bottle). In this case, these objects are spinning objects, thus we used 2D search space definition, which has freedoms along with x and y axis and z position of the object is assumed to be the table height.

The right bottom table shows task relevant visual behavior verification knowledge. Recognizing tea is utilized for pouring tea behavior verification. Recognizing water is applied to verify the open and close tap behaviors. This process is presented in the section IV-C.2.



Fig. 12. daily life support experiments using knowledge based on recognition system.

### C. Evaluation

1) *Robustness*: The demonstration of “The Real-world Information System Project” was successfully performed and covered by major national papers and TVs. It also reported internationally via CNN<sup>1</sup>, USA TODAY<sup>2</sup>. The experiment is repeated more than a dozen times on the day and afterward on demand. Thanks to the robust object recognition using method attention control, visual future prediction, multi-cue integration and visual behavior verification, the task is rarely failed thus we believe the targeted robustness were reached. The rare case of the failure is when there are droplet on the cup or the bottle. It slips when the robot holds them.

Changes of the lighting condition usually affects the recognition result, however our object recognition system is

robust enough not to restrict using flash when taking photos. 3D feature points and 2D edges are robust to the illumination change and we use the HSV color space based histogram matching.

2) *Limitation*: Currently, our system requires knowledge description as shown in the TABLE I and it does not have capability of learning new object and situation. Online acquisition of these knowledge is our current research interest. Our approach is to regard the system described in this paper as a basic function of a humanoid robot. Thus the learning process is able to acquire knowledge based on bootstrap approach by using high level functions presented here. In other words, the result of this research provides the knowledge representation to be learned and online acquisition process is to generate description from an observation or experiences.

3) *Perspective*: Fig.13 shows a multi humanoid daily assistive task. This task is performed in the same environment but two legged humanoid and two wheeled humanoid

<sup>1</sup>Robot serves tea just the way Japanese like it: March 2, 2007

<sup>2</sup><http://www.usatoday.com/tech/news/robotics/2007-02-28-tea-robot.x.htm>

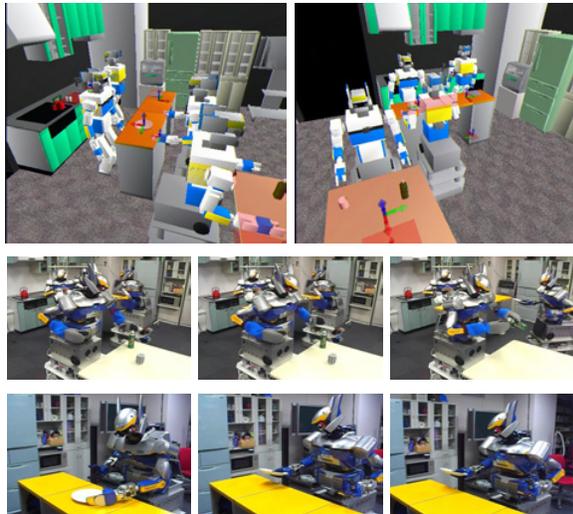


Fig. 13. Multi humanoid experiment in kitchen service task  
 Top row) Knowledge used for the demonstration. Middle row) Humanoid C pouring tea. Bottom row) Humanoid B carrying a dish.

cooperate. These robot treat the dish and the table in addition to the object described in the TABLE I. Thus we defined new behaviors includes “Move to table”, “Hold the dish” and “Place the dish”, describe visual feature knowledge on the table and the dish and new search area on the table. This experiment shows the scalability of our system. Once we describe basic behaviors and objects, it is easy to expand descriptions and realize different task. In fact, the complexity of the system does not increase linearly as number of a robot or behavior increase.

## VII. CONCLUSION

This paper describes daily assistive task experiments that conducting on our HRP2JSK humanoid robot. In order increase robustness, we introduced attention and behavior control method based on visual navigation using task relevant knowledge. The main contribution of this paper is to present the knowledge representation sufficient to perform the humanoid daily assistive task with visual attention and behavior control and demonstrated along with the real humanoid tea serving experiments.

The object recognition method presented here employ the multi cue integrated recognition is currently becomes common technique (for example [16]), however, we have shown that the combination of 3D feature point, color histogram and 2D-3D edge matching are able to cover vision based humanoid behavior generation includes both manipulation and navigation. In this system, we did not integrated visual SLAM [17], [18] for obtaining current location of the robot, since the SLAM provides a geometrical map. However, our system requires relative location from the settled objects as kitchen and the counter bar, since the humanoid manipulation task as “open water tap” and “place cup” are not presented in the world coordinate frame, but described relative to spot knowledge which associated to the fixed object. This we used object recognition method for recognizing. Of course, SLAM

based navigation technique can be integrated to generate collision free path from a spot to another to increase robustness.

## REFERENCES

- [1] H. Inoue, S. Tachi, K. Tanie, K. Yokoi, S. Hirai, H. Hirukawa, K. Hirai, S. Nakayama, K. Sawada, T. Nishiyama, O. Miki, T. Itoko, H. Inaba, and M. Sudo. HRP: Humanoid Robotics Project of MITI. In *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots (Humanoids 2000)*, 2000.
- [2] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura. The intelligent ASIMO: System overview and integration. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02)*, pages 2478–2483, 2002.
- [3] T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann. ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control. In *2006 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)*, pages 169–75, 2006.
- [4] Charles C. Kemp, Aaron Edsinger, and Eduardo Torres-Jara. Challenges for robot manipulation in human environments. *IEEE Robotics & Automation Magazine*, 14(1):20–29, 2007.
- [5] A. Edsinger and C. Kemp. Manipulation in Human Environments. In *2006 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)*, pages 102–109, 2006.
- [6] E. S. Neo, O. Stasse, Y. Kawai, T. Sakaguchi, and K. Yokoi. A unified on-line operation interface for humanoid robots in a partially-unknown environment. In *Proceedings of The 2006 IEEE International Conference on Robotics and Automation*, pages 4437–4439, 2006.
- [7] R. Zollner, T. Asfour, and R. Dillmann. Programming by Demonstration: Dual-Arm Manipulation Tasks for Humanoid. In *Proceedings of the 2004 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS'04)*, pages 479–484, 2004.
- [8] K. Okada, T. Ogura, A. Haneda, J. Fujimoto, F. Gravot, and M. Inaba. Humanoid Motion Generation System on HRP2-JSK for Daily Life Environment. In *2005 IEEE International Conference on Mechatronics and Automation (ICMA05)*, pages 1772–1777, 2005.
- [9] K. Okada, M. Kojima, S. Tokutsu, T. Maki, Y. Mori, and M. Inaba. Multi-cue 3D Object Recognition in Knowledge-based Vision-guided Humanoid Robot System. In *Proceedings of the 2007 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS'07)*, pages 3217–3222, 2007.
- [10] K. Okada, M. Kojima, Y. Sagawa, T. Ichino, K. Sato, and M. Inaba. Vision based behavior verification system of humanoid robot for daily environment tasks. In *2006 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)*, pages 7–12, 2006.
- [11] Genshiro Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, March 1996.
- [12] Michael Isard and Andrew Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [13] Jan Giebel, Dariu Gavrila, and Christoph Schnörr. A bayesian framework for multi-cue 3d object tracking. In *ECCV (4)*, pages 241–252, 2004.
- [14] K. Okada, S. Tokutsu, T. Ogura, M. Kojima, Y. Mori, T. Mak, and M. Inaba. Scenario controller for humanoid using visual verification, task planning and situation reasoning. In *The 10th International Conference on Intelligent Autonomous Systems*, page (to appear), 2008.
- [15] Tomomasa Sato. Real World Informatics Environment System. In *the 9th International Conference on Intelligent Autonomous Systems (IAS-9)*, pages 19–29, 2006.
- [16] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision (ECCV)*, pages 28–39, 2004.
- [17] O. Stasse, A. Davison, R. Sellaouti, and K. Yokoi. Real-time 3D SLAM for Humanoid Robot considering Pattern Generator Information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 348–355, 2006.
- [18] S. Thompson, S. Kagami, and K. Nishiwaki. Localisation for autonomous humanoid navigation. In *Proceedings of 2006 IEEE-RAS International Conference on Humanoid Robots (Humanoids2006)*, pages 13–19, 2006.