Grasp Observation and Reproduction by Humanoid Robots Using Color Camera and 3D Sensor

Tien Cuong KIEU[*1], Kimitoshi YAMAZAKI[*1], Ryo HANAI[*1], Kei OKADA[*1], Masayuki INABA[*1]

***Abstract*— In this paper, we present a system for grasp observation, mapping and execution on humanoid robots to provide intuitive and natural way of communication between humans and robots. This system enables a human user to teach a robot how to grasp an object. The system includes four main components: the hand movement tracking component which provides the approach direction toward the object, the pre-grasp hand pose estimation component which provides the grasp type performed by the human user, the contact points estimation component which provides the position on the object should be grasped, and the grasp mapping and execution component for grasp reproduction on humanoid robots with gripper hands.**

I. INTRODUCTION

Programming robots for new tasks requires many efforts of both programming interface and users. It has been argued that the Programing by Demonstration paradigm can make it easier for unexperienced users to integrate complex tasks in robotic systems [13]. The aim of The Programming by Demonstration is to use natural ways of human-robot interaction where robots can be easily programmed for new tasks by simply observing humans performing the task. To realize that, the first problem to be concerned is understanding human action.

In this research, we design a system that extracts the information needed for one of the most popular action carried out by humans, grasp action. This information is then used to synthesize the motion of robots that do the same action as humans do.

The first thing to be considered is what kind of information needed for robots to carry out grasp action. Besides an object's position and orientation, there are two kinds of parameters needed for a robot to grasp the object. The first one is the hand pose that the robot must do to grasp the object. The second one is the contact position between the robot hand and the object.

For information involving the hand pose, robots need not only to understand human hand pose, but also need to map this hand pose to the category of hand pose that robots can perform. In this research, we consider robots with gripper hands. Therefore, for robots to perform grasp action, there are six categories of hand pose (figure 1). We will present a method of recognizing human hand pose and map it into one of these categories.

[*1] Department of Mechano–Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan. {kieu, yamazaki, hanai, k-okada, inaba}@jsk.t.u-tokyo.ac.jp

For information involving the contact position between a hand and an object, our approach uses data from 3D sensor to extract contact points between the human hand and the object.

Besides components involving grasp observation, another crucial part of our system is the grasp reproduction component, concerning computation of robot grasp and execution of grasp action on humanoid robots based on the information learned in human grasp.

The remain of this paper is organized as follows: Section II introduces a number of research papers related to our work. Section III describes components performing grasp observation. Section IV shows the principle of computing information needed for robot grasp from information learned in human grasp. Section V, VI show the experimental results and state the conclusion of this paper.

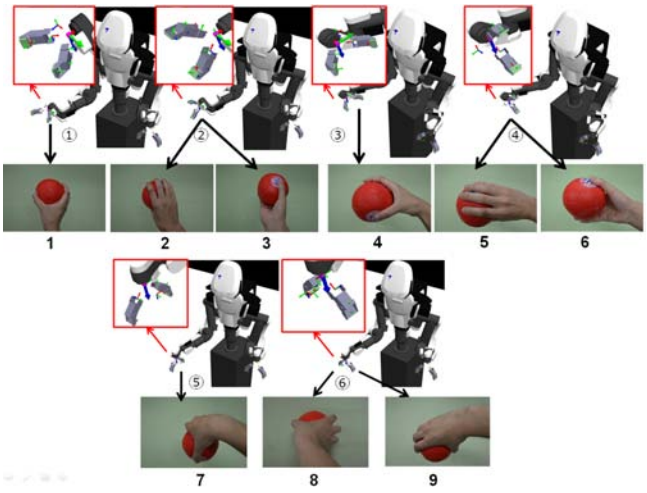

Fig. 1. 6 fudamental types of robot grasp and 9 types of human grasp

II. RELATED WORK

There have been many research papers tackling the problem of teaching robot to grasp by demonstration. In [1], an approach using 3D model for training was presented. The simulation program used Markov Random Fields to classify points on an object. After learning process, points on the object which had high probability of success for grasp were obtained. This approach focused on learning the optimal contact points between robot hands and objects. No vision information or human demonstration was used in the learning process. On the other hand, [2] presented a method of grasp observation and mapping on humanoid

robots. In this case, a robot learned grasp action directly from human's demonstration. The core of this approach was using vision information in the learning process and the problem of recognizing human hand pose while grasping an object was highly focused. The problem involving contact position between hands and objects was not considered. The robot observed human motion and recognized human hand pose, then reproduced all. As a result, the robot was able to grasp an object if this object's position with respect to the robot was the same as the position of the object in the demonstration with respect to the human. Another approach using Neural Network was presented in [3]. A robot learned to grasp an object by itself. Using information from tactile sensors attached on hands, the robot explored its own movement possibilities to interact with objects of different shape, size and material and learned how to grasp them. The method presented also did not use vision information and there was no human's demonstration. One more research which tackled the problem of teaching robot an action by simulation was [4]. In the research, virtual reality was utilized to synthesize robot grasp action. The method required an experienced user to demonstrate typical robot grasps by guiding a robot hand in real time in a virtual environment. [5] had similar approach. The teaching phase mentioned in [5] was executed by human teleoperating a simulated robot to pick up a simulated object. In contrast, [6] presented a method using human's demonstration to teach robots. The system of two cameras was used to track the movement of an object manipulated by a human user. The robot then reproduced this motion. The differences of the approach comparing to other approaches was in the tracking process. Unlike other approaches which tried to track human hand movement, the approach in [6] came up with tracking the movement of the object.

So far, it is able to see that there are two categories of approach when tackling the problem of teaching robot to grasp an object. Approaches in the first category usually try to solve the problem by simulation. Our work falls into the second category, which tries to come up with approaches involving human's demonstration and utilizing of vision information. The system presented in this paper (figure 3) does not only deal with problem of the hand pose for grasp but also consider the contact points between hands and objects. In the grasp reproduction phase, our system can make a robot grasp an object on positions different from the position in the demonstration.

III. GRASP OBSERVATION

A. General Approach

As mentioned in the introduction, we assume that grasp action includes two stages: the approach stage and the final grasp stage. Observation of the whole grasp action involves recognition of grasp type, estimation of approach movement of hand, estimation of contact points between human hands and objects.

In [2], recognition of grasp type is executed in the final grasp stage. On the other hand, our system recognizes grasp type in the approach stage. The reason is that, for many cases, two objects with different shapes, sizes and grasp positions conduce to different kinds of hand pose in the final stage of grasp action, but the hand pose in the approach stage are very similiar (figure 2). Especially, for humanoid robots with gripper hands, we consider that robot grasp consists of six fundamental types (figure 1): three kinds of wrist direction x two kinds of gripper rotation.

So, we consider that recognition of human hand pose in the approach stage, so called pre-grasp hand pose is enough for robots to learn how to grasp an object. There are nine types of human pre-grasp hand pose coresponding to six fundamental types of robot grasp (figure 1). Then, the problem of recognizing human hand pose in the final stage of grasp can be converted to much easier problem of classifying human hand pose in the aprroach stage of grasp into one of these nine types.

For estimation of the approach movement, we do hand movement tracking using 3D sensor mounted on robot head. Estimation of contact points between human hands and objects is also conducted using information given by 3D sensor.

Fig. 2. Hand pose in the approach stage and in the final grasp stage

B. Hand Movement Tracking

As shown in figure 3, our system consists of five components: three components for grasp observation and two components for grasp reproduction.

In order to estimate approach direction toward an object, the human hand movement tracking component (figure 4) is used. The supposed situation is that a human user will grasp an object on a table. 3D sensor gives a point cloud of every thing in the visible area. To extract the point cloud of human hand from this big point cloud, we use color filter. Every point in the big point cloud given by 3D sensor which has skin color is extracted. To reduce noise in case of the back ground or the object has some points with skin-liked color, we do filtering in the folllowing steps.

First, we eliminate unnecessary points by filtering out all points which are far from the observed area. Then, the plane of the table is segmented using RANSAC-based-segmentation algorithm [12]. Now, the remained point cloud includes the object and the human hand when the human

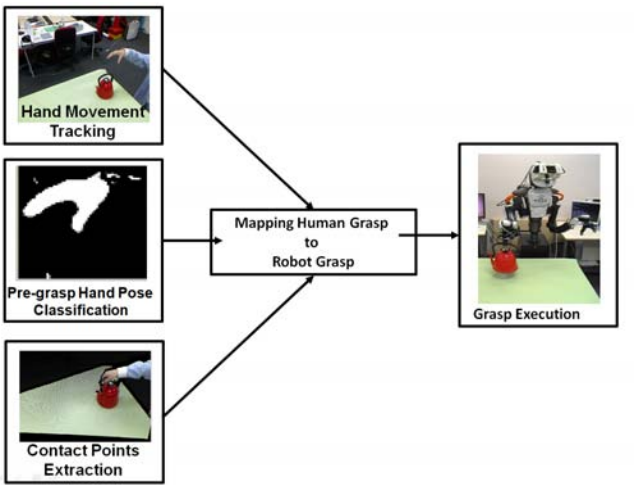

Fig. 3. Overview of the system structure

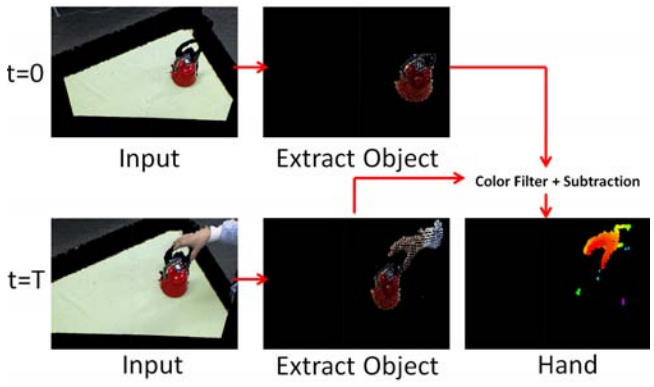

Fig. 4. Hand movement tracking component

grasps the object or only the object if the human still has not grasped the object.

The next step is to extract the human hand from the point cloud of the object and the human hand. To do that, the point cloud of the object should be known. The point cloud of the object is get by taking the point cloud at the time when the human has not grasped the object. Set this time $t = 0$. So, for every time $t = T$, the point cloud of the human hand equals the difference between the point cloud at time $t = 0$ and the point cloud at time $t = T$.

$$P = A - B \tag{1}$$

Where P is the point cloud of the hand only, A is the point cloud of the hand and the object, and B is the point cloud of the object only.

Then, the center of the point cloud of the human hand is calculated. The position of the center of the human hand is recorded for every time $t = T$. When the closest distance between the point cloud of the human hand and the point cloud of the object becomes smaller than a threshold value, i.e, ending point of the approach stage, the process of recording position of the center of the human hand will finish.

C. Pre-grasp Hand Pose Recognition

For the reasons mentioned above, instead of recognizing human hand pose in the final grasp stage, we do this in the aprroach stage. Considering direction and rotation of wrist, human hand pose in the approach stage can be divived into nine fundamental types as shown in figure 1.

The input of the pre-grasp hand pose recognition component is an image of a human hand (figure 5). To get the image of the human hand from the original scene, we do image segmentation with support of 3D information. In Section III.A, we described how to extract the point cloud of the human hand from the large scene. The result of this process, the center of the human hand, is projected into 2D image to get the position of the ROI of the human hand. To define size of the ROI, one more point is projected into 2D image. The distance from this point to the center of the human hand is set up heuristically so that the ROI is able to cover the whole human hand.

Comparing to the method of just extracting human hand from 2D image, this method have benefit that the scale of human hand with respect to the size of ROI image is conservable. Because the ratio between human hand and the distance of two points defining ROI is conservable. The conservability of the scale of human hand in the image is needed because we will use HLAC feature [7], [8] to recognize human hand pose.

The image of human hand is then filtered by skin color again to get the binary image of human hand. HLAC feature of the binary image is calculated. HLAC feature is used to find the most similar image in the data base. Searching algorithm is Nearest Neighbor Search algorithm.

To reduce the number of nodes searched and increase recognition accuracy, we use approach direction estimated in Section III.B as constraint condition for searching. The images of human hand in data base are grouped into three groups corresponding to approach direction x, or y or z in the world coordinate system. For instance, if the approach direction is x, we only search similiar image in the corresponding group.

D. Contact Points Estimation

For grasping an object, especially an object used as a tool, the position on the object should be grasped is very important. Because unlike the other objects, an object used as a tool need to be grasped at the special position, so that human can use it. The position on the object should be grasped can be understood by observing the contact position between human hand and object in the final grasp stage.

To obtain contact position between human hand and object, we use 3D sensor. 3D sensor is able to get a point cloud

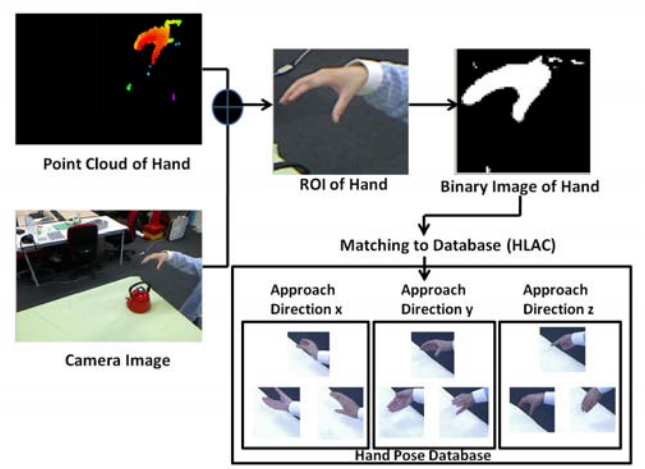

Fig. 5. Recognition of pre-grasp hand pose

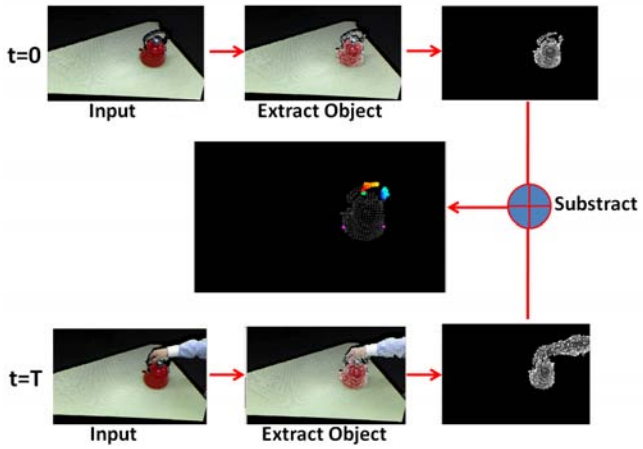

Fig. 6. Estimation of candidates of contact points

of object and human hand and as a result, the contact points between human hand and object can be computed.

First, as mentioned in Section III.A, before the human grasps the object, there is only object in the scene. We extract the point cloud of the object from the original scene and save it as a reference point cloud. Then, when human begins grasp action, for every time $t = T$, the point cloud including human hand and grasp object is extracted from the scene. The difference between this point cloud and the reference point cloud is calculated. The result is set of points that are in the reference point cloud but not in the point cloud of frame $t = T$. This is the point cloud of the part on object that is occluded by human hand when human grasps the object. So, this point cloud covers the contact position betweetn human hand and object. In the other words, this point cloud is set of candidates for contact points between human hand and object.

Next step is to estimate contact points from these candidates. The estimation is executed using results in hand movement tracking step and pre-grasp hand pose recognition step. The details will be discribed in Section IV.

IV. GRASP REPRODUCTION

By observing the human user's grasp action, the robot acquires a lot of information. For the robot to reproduce grasp action, the next step is to convert this information into robot's grasp action. Generally, for a humanoid robot to perform grasp action, there are three kinds of information needed. The first one is the configuration of robot hand in the approach stage, i.e, the coordinate, orientation of robot hand. This information is shown by approach frame (figure 7). The second and third information involves the configuration of robot hand in the final grasp stage. These information are defined by a target frame and grasp width. In the following, we will describe how to compute these three parameters from information get in the observation.

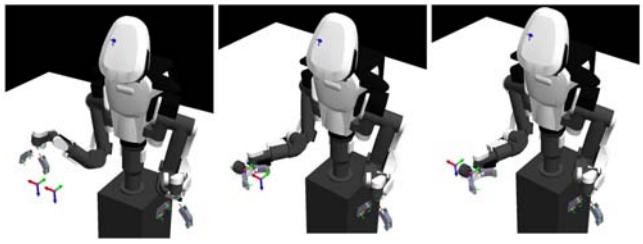

Fig. 7. Robot motion for grasp an object

A. Computation of approach frame, target frame and grasp width from observed information

The first step is to compute the orientation of robot hand. The orientation of robot hand in the aprroach stage is expressed by the x, y, z axis of the approach frame. The x axis of the approach frame is computed from the information get in human hand movement tracking component.

The human hand tracking component records the movement of human hand while grasping an object. As mentioned in Section III, this information is stored in form of an array of coordinate of the center of human hand. So, in fact, every element in this array is a point in 3D space. We do line fitting to the set of these points and the obtained line is the approach direction toward object in grasp action. The x axis in the approach frame is this aprroach direction.

The orientation of y axis and z axis is define by the result of pre-grasp hand pose recognition. The figure 8 shows the relation between the orientation of y axis and z axis of robot hand with respect to pre-grasp hand pose. Finally, the position of robot hand frame in the approach stage is the last point in the array of coordinate of the center of human hand get by the human hand tracking component.

For the target frame, the orientation of x,y,z axis is the same as in the approach frame. However, the position of the target frame is concerned to the contact points between hand and object. So, to define position of the target frame, it is needed to compute the contact points between hand and

object.

The result from the contact points estimation component is set of candidate points for contact points. From these candidate, two points, the contact points between robot gripper and object will be computed. Let us consider the set of these candidate points as a surface (curved or plane). The contact points are tangent points of robot gripper and this surface. These tangent points are determined by the direction of gripper and orientation of wrist (figure 8). Let us define x_{min} as the minimum of x component, x_{max} as the maximum of x component for all points in the set of contact point candidates. Similarly for y_{min}, y_{max}, z_{min}, z_{max}.

Corresponding to six fundamental types of robot grasp (figure 1), the contact points, the position of target frame and the grasp width are computed as follows.

TABLE I

COMPUTATION OF CONTACT POINTS, TARGET FRAME AND GRASP WIDTH

Grasp Type	C_{1p}	C_{2p}	T_p	W
(1)	(x_{max},y_{min}, z_{ave})	(x_{max},y_{max},) z_{ave}	(x_{max},y_{ave}, z_{ave})	y_{max}-y_{min}
(2)	(x_{max},y_{ave}, z_{min})	(x_{max},y_{ave},) z_{max}	(x_{max},y_{ave}, z_{ave})	z_{max}-z_{min}
(3)	(x_{min},y_{max}, z_{ave})	(x_{max},y_{max},) z_{ave}	(x_{ave},y_{max}, z_{max})	x_{max}-x_{min}
(4)	(x_{ave},y_{max}, z_{min})	(x_{ave},y_{max},) z_{max}	(x_{ave},y_{max}, z_{ave})	z_{max}-z_{min}
(5)	(x_{ave},y_{min}, z_{min})	(x_{ave},y_{max},) z_{min}	(x_{ave},y_{ave}, z_{min})	y_{max}-y_{min}
(6)	(x_{min},y_{ave}, z_{min})	(x_{max},y_{ave},) z_{min}	(x_{ave},y_{ave}, z_{min})	x_{max}-x_{min}

Where C_{1p} is coordinate of the first contact point, C_{2p} is coordinate of the second contact point, T_p is coordinate of the target frame, W is the grasp width. For x, x_{ave} = (x_{min} + x_{max})/2. Similarly for y and z.

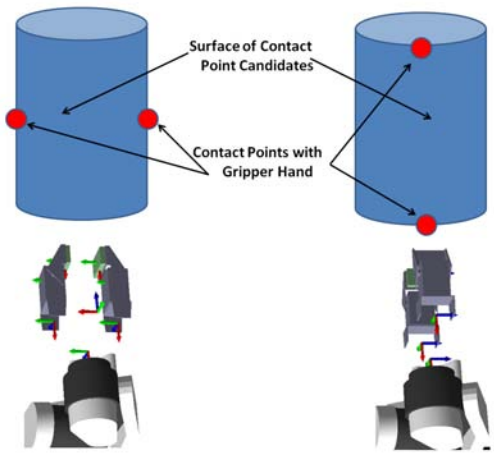

Fig. 8. Computation of contact points between robot hand and object

V. EXPERIMENTS

A. Experiment Setup

The humanoid robot HIRO was used as the experimental platform in this work. From the kinemactics point of view, the humanoid robot HIRO consisted of seven components: head, right arm, left arm, right hand, left hand, torso and base. The head had 2 DoF and was equiped with three cameras and one Kinect. The right and the left hand of HIRO were also equiped with one hand-camera for each hand. In this work, we only used 3D sensor and color camera in the Kinect to get vision information.

For the experiments, we used objects with AR-markers attached on. The experiment setup stipulated that the demonstration of grasp action was performed by a human user standing beside the robot.The experiments were conducted with the following objects: box, dish, plastic bottle, mug cup and kettle.

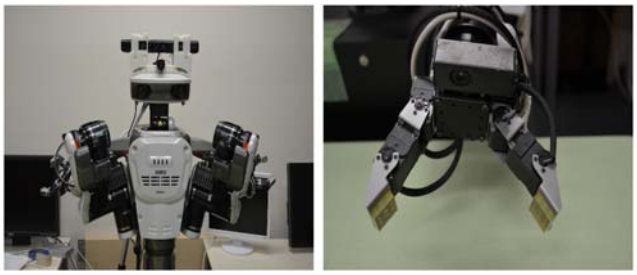

Fig. 9. Left: The humanoid robot HIRO. Right: HIRO's hand.

B. Experimental Results

As depicted in figure 10 and 11, the robot successfully imitated the grasp action demonstrated by the human user. The experimental objects included: box, dish, plastic bottle, mug cup and kettle. The first experiment was conducted with three objects which have simple shape: box, dish and plastic bottle. The grasp action of these three objects covered all fundamental types of grasp shown in figure 1. The grasp action of box corresponded to grasp type 7,8 and 9 of human grasp (grasp type 5 and 6 of robot grasp). The grasp action of dish could be represented for grasp type 2,3,5,6 of human grasp (grasp type 2,4 of robot grasp). The grasp action of plastic bottle corresponded to grasp type 1,4 of human grasp (also grasp type 1,3 of robot grasp). After observing human demonstration, the robot successfully grasped all three objects. The robot could also adapt to grasp the objects even in the case the objects were located in positions and orientations different with positions and orientations in the human demonstration.

The second experiment was conducted with a kettle and a mug cup. The grasp action of kettle represented for grasp type 7 of human grasp (grasp type 5 of robot grasp) and the grasp action of mug cup represented for grasp type 4 of human grasp (grasp type 3 of robot grasp). The difficulty of grasping ketttle and mug cup was that the two objects have

to be grasped on special positions because of its usage. The robot successfully grasped the two objects in the same way as the human user did.

The experimental results show that with our system, a human user can teach a robot all fudamental types of gripper-handed robot's grasp just by demonstration. Moreover, the success of teaching robot to grasp the kettle and the mug cup proves that our system is effective to teach robots to grasp objects with complex shape and objects which have specical constraints of grasp defined by its usage in human society.

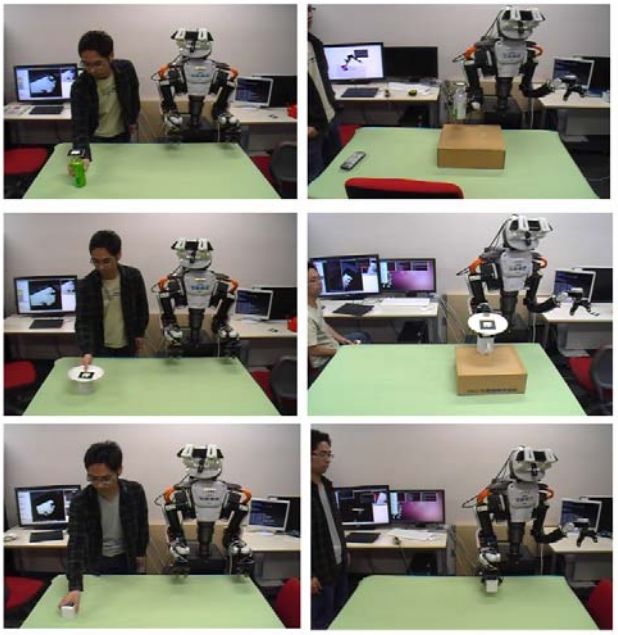

Fig. 10. Plastic bottle, dish and box grasp

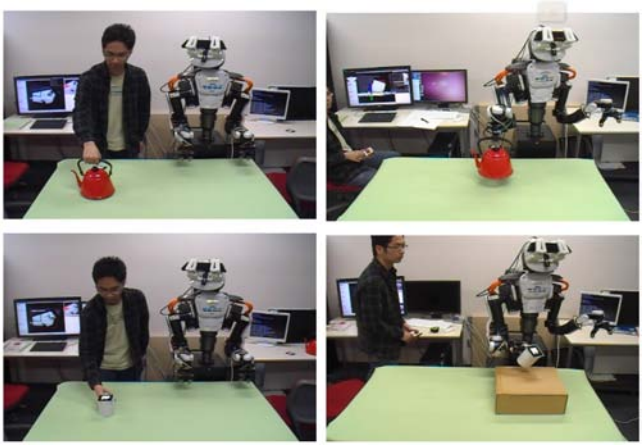

Fig. 11. Kettle and mug cup grasp

VI. CONCLUSION

In this paper, we presented a system for observing human grasp action, mapping and reproduction on a humanoid robot. Human grasp action is considered to consist of two stages: the approach stage and the final grasp stage. The observation of the approach stage provides robots the information involving the approach movement of hand and the pre-grasp hand pose to grasp an object. In the final grasp stage, information involving contact position between hands and objects is observed. From these kinds of information, the mapping from human grasp and robot grasp is executed and robots reproduces the demonstrated grasp action. Especially, the observation of contact positon between human hand and object makes robots be able to grasp objects at special positions depending on its usage. In general, the proposed system is one step to make the human-robot interaction more friendly and make the usage of robot easier for normal users.

REFERENCES

[1] A. Boularias, O. Kroemer and J. Peters, "Learning Robot Grasping from 3-D Images with Markov Random Fields", *in Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'11)*, San Francisco, CA, USA, 2011.

[2] Martin Do, J. Romero, H. Kjellstrom, P. Azad, T. Asfour, D. Kragic and R. Dillmann, " Grasp recognition and mapping on humanoid robots", *in IEEE-RAS International Conference on Humanoid Robots*, 2009.

[3] G. Gomez, A. Hernandez, P. E. Hotz and R. Pfeifer, "An adaptive learning mechanism for teaching a robot to grasp", *in International Symposium on Adaptive Motion of Animals and Machines (AMAM 2005)*, Illmenau, Germany, 2005.

[4] J. Aleotti and S. Casalli, "Robot Grasp Synthesis from Virtual Demonstration and Topology-Preserving Environment Reconstruction", *in Intelligent Robots and Systems, IROS 2007*, San Diego, CA, 2007.

[5] K. Hsiao and T. Lozano-Perez, "Imitation Learning of Whole-Body Grasps ", *in International Conference on Intelligent Robots and Systems, 2006*, Beijing, 2006.

[6] Y. Maeda, N. Ishido, K. Haruka and T. Arai, "Teaching of Grasp/Graspless Manipulation for Industrial Robots by Human Demonstration", *in Intl. Conference on Intelligent Robots and Systems, 2002*, Lausanne, 2002.

[7] N. Otsu, "A new scheme for practical, flexible, and intelligent vision systems", *in Proc. IAPR Workshop on Computer Vision*, pp.431-435, 1988.

[8] T. Kurita and S. Hayamizu, "Gesture recognition using HLAC features of PARCOR images and HMM based recognizer", *in Proc. Intl. Conference on Automatic Face and Gesture Recognition, 1998*, Nara, 1998.

[9] H. Kjellstrom, J. Romero and D. Kragic, "Visual recognition of grasps for human-to-robot mapping", *in Intl. Conference on Intelligent Robots and Systems, 2008, IROS 2008*, Nice, 2008.

[10] M. Ester, H. Kriegel, J. S and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise ", *in Intl. Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996.

[11] T. Feix, O. Bock, H. Gmbh, H. Schimiedmayer, J. Romero and D. Kragic, "A comprehensive grasp taxonomy", *in Robotics, Science and Systems Conference: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation* , 2009.

[12] R. Schnabel, R. Wahl and R. Klein, "Efficient RANSAC for Point-Cloud Shape Detection", *Computer Graphics Forum*, vol. 26, No. 2. (June 2007), pp. 214-226.

[13] S. Ekvall, "Robot task learning from human demonstration", Ph.D. dissertation, KTH, Stockholm, Sweden, 2007.