Learning to Grasp with Parental Scaffolding

Emre Ugur[1,2], Hande Celikkanat[2,3], Erol Şahin[3], Yukie Nagai[4] and Erhan Oztop[1,2,5,6]

[1] Advanced ICT, National Institute of Information and Communication Technology, Kyoto, Japan
[2] Cognitive Mechanisms Labs., Advanced Telecommunications Institute International, Kyoto, Japan
[3] KOVAN Research Lab., Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
[4] Graduate School of Engineering, Osaka University, Osaka, Japan
[5] School of Engineering Science, Osaka University, Osaka, Japan
[6] Computer Science Dept., Ozyegin University, Istanbul, Turkey
Emails: emre@atr.jp, hande@ceng.metu.edu.tr, erol@ceng.metu.edu.tr, yukie@ams.eng.osaka-u.ac.jp, erhan@atr.jp

***Abstract*—Parental scaffolding is an important mechanism utilized by infants during their development. Infants, for example, pay stronger attention to the features of objects highlighted by parents and learn the way of manipulating an object while being supported by parents. In this paper, a robot with the basic ability of reaching for an object, closing fingers and lifting its hand lacks knowledge of which parts of the object affords grasping, and in which hand orientation should the object be grasped. During reach and grasp attempts, the movement of the robot hand is modified by the human caregiver's physical interaction to enable successful grasping. The object regions that the robot fingers contact first are detected and stored as potential graspable object regions along with the trajectory of the hand. In the experiments, we showed that although the human caregiver did not directly show the graspable regions, the robot was able to find regions such as handles of the mugs after its action execution was partially guided by the human. Later, this experience was used to find graspable regions of never seen objects. At the end, the robot was able to grasp objects based on the position of the graspable part and stored action execution trajectories.**

I. Introduction

Scaffolding in developmental psychology refers to the support from an (adult) caregiver in order to speed up a child's skill and knowledge acquisition [1]. This support can take various forms, including the attraction and maintenance of the child's attention on relevant items, the shaping of the environment in order to ease the task (such as positioning and orienting the child so as to limit its degree of freedom), signalling the important features or subgoals of the task, or providing feedback and reinforcement [2]. When the task gets out of hand, puzzling the child, the caregivers step in as the "trouble-shooter" [3]. They interfere at different steps of the task, initially demonstrating the goal and drawing attention to task-relevant features, then "embodying" the child to co-achieve the goal. Throughout the process, they let the child have proprioceptive, force, tactile, visual, and auditory feedback, until the goal is achieved. Moreover, it is not only the adults who are interested in this kind of interaction. By acquiring the ability for joint attention, the children become aware of the caregiver as a "helper", and begin "asking" for help when faced a difficult task by displaying significant communicative gestures. From the age of 9-months, when the joint attention mechanisms begin to emerge, to 18-months, infants use more and more of these communicative signals [4]. These interactions have a purpose. Infants can exhibit certain skills in game-contexts together with their mothers far before they can perform them in isolated cognitive tests [5].

The idea of parental scaffolding seems especially inspiring from a robotics point of view. The idea has been exploited in various studies with different viewpoints, such as for better communication between humans and robots [6], or as a grounding principle for lifelong developing of robots "at home" [7]. It has been shown that caregivers tend to modify their motions when teaching a task to a child. Analogous to "motherese", [8] call these motions of higher interactiveness, enthusiasm, proximity, range of motion, repetitiveness and simplicity, as "motionese". In a similar line, [9] reveals a significant amount of bottom-up saliency features in infant-directed interaction versus adult-directed interaction. Motivated by these findings, [10] develops a bottom-up architecture for robot-infants, who, like human-children, are equipped with minimal a-priori information, and therefore in need of depending mainly on bottom-up signals as much as possible. Interestingly, such infant-robots are also found to motivate humans to use motionese as if they were dealt like human children. Due to the robots' limited attention mechanisms, humans try to carefully teach a task for example by approaching to the robots and introducing the object closely to their attention, sometimes shaking it, amplifying their movements and making pauses. Evidently, these are also widely used tactics in parent-infant communication.

Another issue in robot imitation is the inability of the robot to understand "what" to imitate. In particular, there are goal-oriented tasks, where any means to achieve the goal are acceptable, versus means-oriented tasks, where the motion itself is equally important. This is also a problem faced by human infants. Here parents again come to rescue by signalling the important features [11]. In a goal-oriented task, they emphasize initial and final states, as well as important subgoals, by taking long pauses. Conversely, in a means-oriented task, they emphasize the movement itself by adding additional movements to the object, for instance by shaking it.

Scaffolding has also been used as a means of "correcting" the robot's experiences and letting it learn the "right" way: [7]

demonstrates the scaffolding of the environment itself as one way of reducing complexity. The robot is expected to perform pre-taught behaviors, such as wall-following, by matching its current sensory values to previously memorized instances. The most "similar" previous instance is decided, however, by giving more weight to features with higher information gain for the specific task. Here comes in the human teacher, who modifies the environment in the learning phase, by reducing variations in irrelevant features. These features, having always constant values, will not affect the robot's behavior later on. [12] takes a more direct approach, by correcting the very movement of the robot. The robot is allowed to learn a behavior from demonstrations by a teacher. Once it derives a policy, the human will help it correct its movement by online tactile feedback from sensors attached to its wrist.

In [13], social aspects of scaffolding has been explored. The robot tries to increase its own skills through its intrinsic drives of novelty and mastery, but a human can also draw its attention, guide it by suggesting actions, point out possible goal states and structure the environment so that certain components become more salient. These signals act as external cues triggering the robot's reinforcement learning system.

In this paper, we propose and implement a robotic framework where a human caregiver speeds up robot's affordance [14] acquisition through parental scaffolding. In particular, the robot learns how to detect graspable object regions and how to act upon these detected affordances in order to achieve the goal of grasping and lifting. In our previous studies [15], [16], a similar anthropomorphic robot self-discovered simple object affordances by learning the relations between objects, effects and behavior parameters. However, the behavior parameter space is very large in grasping with a dexterous robot hand, and many different parts of complex objects can provide graspability. In the current study, a human caregiver speeds up robot learning by physically modifying robot's built-in *reach-grasp-lift* behavior execution trajectory. While being guided by the human, the robot first detects the 'first-contact' points its finger made with the objects, and stores the collection of these points as graspable regions if the object is lifted successfully. Later, it builds up simple classifiers using these experienced contact regions and use these classifiers to detect graspable regions on novel objects. At the end, the robot hand was shown to lift an object in different orientations by selecting one of the experienced trajectories.

II. Scaffolding Framework

In our parental scaffolding framework, the robot has a default *reach-grasp-lift* action where the object is detected by robot's perceptual system, a reach trajectory is computed based on robot's arm kinematics and object center, and the robot fingers are closed they are nearby to the object. The robot has no initial knowledge about graspability of the objects. Different objects can be grasped from different parts with different hand orientations, so *reach to object center* execution should be modified by the human teacher during trajectory execution. Thus, the initial trajectory is modified by incorporating the force applied to the robot hand by the human during the course of the action. The points on the object where fingers make first contact are stored as potential grasp affording parts. After hand closure is completed, the robot lifts its hand and checks whether the object is lifted or not by searching table surface with its perceptual system again and through force sensor measurements. This online, human modified *reach-grasp-lift* action is repeated many times with different object configurations and graspable parts of the objects are discovered by the robot. Below, the tools and methods to realize this framework will be detailed.

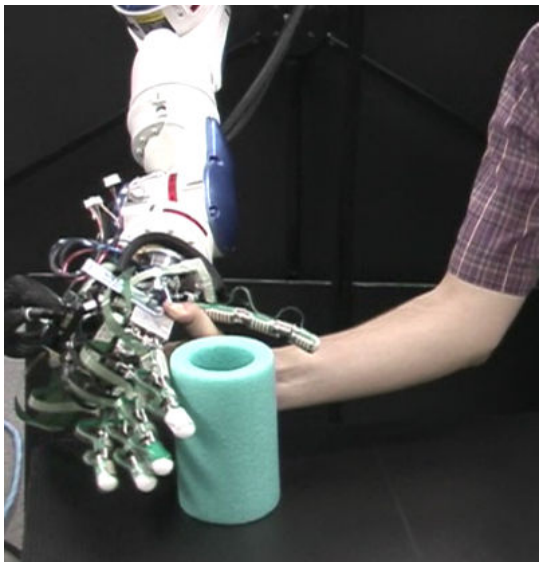

Fig. 1. 7 DoF robot arm, 16 DoF robot hand, table, and a sample object is shown. The range camera is placed on the top-right and not visible. The human teacher can change the default trajectory of the robot to enable grasping thanks to the force/torque sensor that is placed between robot arm and hand.

A. Robot platform

An anthropomorphic robotic system equipped with a range camera is used as the experimental platform. This system uses a 7 DoF Motoman robot arm, that is placed on a vertical bar similar to human arm as shown in Figure 1. A five fingered 16 DoF Gifu robot hand is mounted on the arm to enable manipulation. The maximum length of Motoman arm and Gifu hand is 123 cm and 23 cm, respectively. For environment perception, an infrared range camera (SwissRanger SR-4000), with 176x144 pixel array, 0.23° angular resolution and 1 cm distance accuracy is used. The range camera is calibrated in the coordinate space of the robot hand by computing the transformation matrix between positions extracted kinematically and perceived from range camera. Then, the environment is represented as point cloud in this space. To control robot arm, a 6 DoF Nitta force/moment sensor is mounted between the hand and the wrist of the robot arm.

B. Force-based human-robot interaction

We wanted to be able to guide the robot similar to a caregiver guiding an infant's movement. It is natural for a

human parent to hold the infant's hand, and position it in the space to help with a grasp. A similar intuitive effect can be obtained by attaching a force sensor to right below the hand, at the wrist position. The desired effect is holding the wrist of the robot, and moving the 7-DoF arm in the 6-DoF Cartesian space freely. This requires the force sensed by the robot to be converted into joint displacements.

1) Mapping force and moments to joint displacements: The force sensor continuously outputs two vectors, $\mathbf{f_{app}}$ and $\mathbf{m_{app}}$, the current applied force and moment vectors, respectively. These vectors are first converted to desired position and orientation changes, which are, in turn converted to desired changes in the joint angles of the robot using inverse kinematics.

Initially, $\mathbf{f_{app}}$ is in robot's end-effector frame, since force values are read from a sensor fixed to the robot's wrist. The end-effector frame moves continuously (together with the wrist) with respect to the global frame. However, the human instructor would expect to see the effect of his/her feedback in the global frame. Therefore, the force vector should be represented in the global frame. We identify the robot end-effector frame with two orthogonal vectors $\mathbf{h_y}$, the palm normal and $\mathbf{h_z}$, the middle finger direction when the finger is fully extended ($\mathbf{h_x}$ is the cross product of the two). The force vector in the global reference frame is given by $\mathbf{f_{glob}} = [\mathbf{h_x}\ \mathbf{h_y}\ \mathbf{h_z}]\,\mathbf{f_{app}}$. The desired change in the end-effector position is simply taken as the scaled version of the force represented in the global reference frame i.e. $\Delta\mathbf{p} \propto \mathbf{f_{glob}}$. Similarly, the desired orientation change implied by the human interaction is obtained by considering the scaled moment values as the desired rotations around the end-effector reference frame axes.

$$\Delta\theta = (\Delta\theta_x, \Delta\theta_y, \Delta\theta_z)^T \propto \mathbf{m}_{app}$$

From these angles, the desired (infinitesimal) rotation matrix is obtained with $\mathbf{R} = \mathbf{R}_x(\Delta\theta_x)\mathbf{R}_y(\Delta\theta_y)\mathbf{R}_z(\Delta\theta_z)$ where $\mathbf{R}_x$, $\mathbf{R}_y$ and $\mathbf{R}_z$ represent the rotation matrices around x, y, and z axis, respectively. Given the current end-effector frame $[\mathbf{h_x}\ \mathbf{h_y}\ \mathbf{h_z}]$, the desired end-effector frame is given by

$$\left[\mathbf{h_x^{des}}\ \mathbf{h_y^{des}}\ \mathbf{h_z^{des}}\right] = \mathbf{R}\,[\mathbf{h_x}\ \mathbf{h_y}\ \mathbf{h_z}]$$

which allows the computation of the desired change in the finger direction as $\Delta\mathbf{h_z} = \mathbf{h_z^{des}} - \mathbf{h_z}$.

Having calculated the desired changes in the position and finger direction, these values are converted to desired joint angle changes using the Jacobian of the position and finger direction vector.

$$\Delta\mathbf{q} \propto \mathbf{J^T}\,[\Delta\mathbf{p}\ \ \Delta\mathbf{h_z}]'$$

where $\Delta_\mathbf{q}$ is the desired joint angle change, and the the Jacobian is obtained using the forward kinematics of the robot; i.e.

$$\mathbf{J}\,(\mathbf{q}) = [\partial\mathbf{p}/\partial\mathbf{q}\ \ \partial\mathbf{h_z}/\partial\mathbf{q}]'$$

Since the rotation axis of the last joint of the robot coincides with $\mathbf{h_z}$, for the desired infinitesimal rotation for this joint no inverse kinematics is necessary as this can be obtained directly from the moment reading around this axis (i.e. $\Delta\theta_z$).

2) Gravity Compensation: As the force sensor is connected to the robot hand which has a non-negligible weight, the sensor reports non-negligible force and moment measurements even there is no external force acting on the robot hand. Moreover, these values depend on the joint position and velocities of the robot. If one assumes low velocities, only the orientation of the hand in space can be considered. In principal, this should allow one to obtain equations for the force and moments that will be exerted on the sensor depending on the orientation of the hand. However, due to un-modeled dynamical effects such as the movement of the heavy cables connected to hand we decided to apply a learning approach. We first collected data by having the robot systematically scan a range of orientations ψ_i and record the measured force, $\mathbf{f_i}$ and moment $\mathbf{m_i}$ from the sensor. Orientations are represented as rotation matrices; the force and moment measurements are stored as raw values reported by the force sensor (i.e. three dimensional vectors with respect to the local coordinate frame of the force sensor). The goal is to predict force and moment values given an orientation of the robot hand. For this regression problem, we formed an input data matrix $\mathbf{X}$ where each row corresponds to the elements of the rotation matrix corresponding to the orientation ψ_i. The target matrices $\mathbf{Y_f}$ and $\mathbf{Y_m}$ are formed by the force ($\mathbf{f_i}$) and moment ($\mathbf{m_i}$) vectors, respectively. Using the least squares fit we obtain weight matrices as follow.

$$\mathbf{W_f} = \mathbf{X}^+\mathbf{Y_f} \quad \mathbf{W_m} = \mathbf{X}^+\mathbf{Y_m} \tag{1}$$

With these weight matrices, we can predict the force and moments that will be experienced by the force sensor at a given orientation (by flatting the rotation matrix into a 9 dimensional vector and premultiplying it with the weight matrix). Let these predictors be denoted as

$$\hat{\mathbf{f}}\,(\mathbf{R}) = \left(\hat{f}^x\,(\mathbf{R}), \hat{f}^y\,(\mathbf{R}), \hat{f}^z\,(\mathbf{R})\right)$$
$$\hat{\mathbf{m}}\,(\mathbf{R}) = (\hat{m}^x\,(\mathbf{R}), \hat{m}^y\,(\mathbf{R}), \hat{m}^z\,(\mathbf{R}))$$

Then for quantifying the fit error on test data we define the following relative errors, e_j^{Fx} and e_j^{Mx} with respect to mean force or moment experienced during the testing (e_j^{Fy}, e_j^{Fz}, e_j^{My} and e_j^{Mz} are also defined similarly).

$$e_j^{Fx} = \frac{\left|f_j^x - \hat{f}^x(\mathbf{R_j})\right|}{\mu_{\left|f_k^x\right|}} \quad , \quad e_j^{Mx} = \frac{\left|m_j^x - \hat{m}^x(\mathbf{R_j})\right|}{\mu_{\left|m_k^x\right|}}$$

The data collection for training and testing is conducted as follows. The last three joint angles of the robot are used to obtain a range of orientations.

Starting from the initial joint angle configuration:

$$[q_1\ q_2\ q_3\ q_4\ q_5\ q_6\ q_7]^T = [90^\circ\ 0\ 0\ \text{-}90^\circ\ 0\ 0\ \text{-}80^\circ]^T$$

q_5 and q_6 are varied in the range $[-60^\circ, +60^\circ]$, and q_7 in $[-180^\circ, +180^\circ]$, with 10° intervals, and independently from each other. Sampling with these angles enables the robot to cover the hand orientation space of the grasp executions. The arm is moved to each sample point, and the experienced force and moment values at the configuration are saved.

TABLE I

THE MEAN AND STANDARD DEVIATION OF THE RELATIVE ERRORS MADE ON THE TEST SET FOR FORCE AND MOMENT PREDICTIONS (TWO-FOLD CROSS VALIDATION, 500 REPEATS) ARE SHOWN AS PERCENTAGES.

	$\mu_{E^{force}}$	$\sigma_{E_j^{force}}$	$\mu_{E^{mom}}$	$\sigma_{E^{mom}}$
X	1.88	0.015	2.92	0.025
Y	3.21	0.023	2.86	0.026
Z	1.87	0.013	6.07	0.055

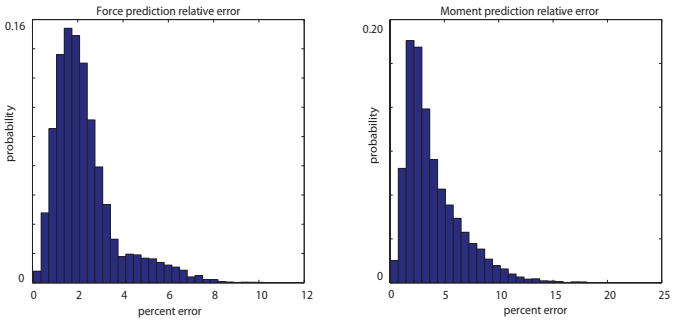

Fig. 2. The normalized histograms of the axis-averaged relative errors for force and moment predictions (i.e. histograms of E^F and E^M).

The prediction accuracy obtained with the collected data is assessed by randomly dividing the data into training and test set. Table I shows the relative errors obtained for each axis as percentages. To verify the performance that we could expect from the prediction we also look at distribution of the errors. For this we define relative errors averaged over the axes as

$$E_j^F = \frac{1}{3}\left(e_j^{Fx} + e_j^{Fy} + e_j^{Fz}\right)$$

$$E_j^M = \frac{1}{3}\left(e_j^{Mx} + e_j^{My} + e_j^{Mz}\right)$$

and compute the error histograms. The histograms obtained are shown in Fig. 2.

C. Robot perception

The robot uses range camera to detect the object on the table. The detected object is represented as a 3D point cloud and from this point cloud various features such as local distance histograms as detailed in the next subsection are computed. Furthermore, the distance between any point on the object and hand fingers is computed by comparing the 3D position of that point measured by range camera and 3D position of the finger computed by forward kinematics based on arm and hand angles. As a result the robot can close its hand when the fingers are nearby to the object or can detect the object points that are in contact with fingers.

The first step of pre-processing is to filter out the pixels whose confidence values reported by range camera are below an empirically selected threshold value. The robot's workspace consists of a black table, so region of interest is defined as the volume over the table, and black pixels are filtered out as the range readings from black surfaces are noisy. As a result, the remaining pixels of the range image are assumed to belong to one or more objects. These objects are segmented by the Connected Component Labeling algorithm which differentiates object regions that are spatially separated by a preset threshold value (2 cm in the current implementation). In order to reduce the effect of camera noise, the pixels at the boundary of the object are removed, and Median and Gaussian filters with 5x5 window sizes are applied. The detected objects on the range image of a sample setup is shown at the top row in Figure 3.

The robot detects the 'first-touch' points on the object surface by checking the distance between each object point and finger joint in each timestep during reach action execution. If the number of object points close to one of the finger joints is larger than a threshold for a certain duration, then those points are stored as grasp points. In our setting, the minimum number of contact points is set as 15 pixels on the range image, the duration is set as 300 msec and closeness threshold is set as 2 cm. Figure 3, third row shows only the detected objects as grayscale and the contact pixels in red color. The bottom row shows the range image at the end of each grasp and lift execution. This simple thresholding may find incorrect touch points in various situations. However, we argue that generalization from multiple interactions and caregiver's own development in scaffolding would solve this problem.

D. A simple classifier based on distance histograms

A successful grasp mediated by caregiver provides very important cues for the further development of the perception system of the agent. To show this, we utilized a series of percepts when the robot touched and subsequently grasped from a part an object to construct a simple (grasp) affordance detector. In this subsection, we will describe our method over the example of handle-grasping to ease understanding, but this method can be used for any graspable part of the object and is not limited to encode only handles. Instead of a pure template based search, we opt to use a feature that will capture the relative property of the grasped-part with regards to the totality of the object. Primate brain is endowed with several shape and topological feature detectors, for example some of which are tuned for rings and hollow objects [17], [18]. Inspired from this, we propose a metric that captures the distribution of three dimensional points (i.e. voxels) that make up a given object. We propose that each voxel is identified by the distribution of its distances from the neighboring voxels that make up the object. This distribution changes smoothly as one moves smoothly on the surface of the object, and is invariant of orientation changes. Our idea was to develop a classifier based on this metric, with the intuition that the handle voxels would have similar distance distributions that are significantly different from the body voxel distributions. For the handle voxels found by interacting with the object are used to construct a distance distribution, p_H. Likewise the rest of the object points is used to construct a distribution representing the non-handle points, p_B. At a later time when the robot faces a novel object it computes a distance distribution for each point on the object and compares it with p_H and p_B, and decides

what points can be used as handles.

Let H and B the set voxels from the handle and the body of the object at a given interaction. We define δ neighbor distance function to operate on a voxel $\mathbf{x}$ (three dimensional vector) and a set of voxels, Y with

$$\|\mathbf{x}, Y\|_\delta = \{\|\mathbf{x}-\mathbf{y}\| : \mathbf{y} \in Y \wedge \|\mathbf{x}-\mathbf{y}\| < \delta\}$$

With this we can compactly define these two distance sets for the handle set and the body set as:

$$\Omega_H = \{\|\mathbf{h_i}, B\|_\delta : \mathbf{h_i} \in H\}$$

$$\Omega_B = \{\|\mathbf{b_i}, H\|_\delta : \mathbf{b_i} \in B\}$$

Taking $hist()$ as an operator to return a normalized histogram of a given set we obtain these two probability density estimates for handle and body voxels, respectively:

$$p_H \sim hist\left(\Omega_{H_1} \cup \Omega_{H_1} ... \Omega_{H_N}\right) \quad (2)$$

$$p_B \sim hist\left(\Omega_{B_1} \cup \Omega_{B_1} ... \Omega_{B_N}\right)$$

To determine whether a given voxel, v belongs to handle or not we compute the distance distribution $p_{\{v\}} \sim hist(\Omega_{\{v\}})$ and compare it to p_H and p_B using the cumulative distribution. The voxel v is deemed a handle voxel if

$$\max \int_0^s p_H - p_{\{v\}} < \max \int_0^s p_B - p_{\{v\}}$$

otherwise it is deemed a body voxel, where the max runs over $s < 1$.

III. Experiments

In this section, we first described the experimental setup where the robot arm is guided by the human caregiver during *reach-grasp-lift* execution, and we showed the detected 'first-touch' points indeed corresponded to graspable parts of the objects. Then, the classifier (Section II-D) that differentiates whether a voxel is graspable or not was trained using this first-touch points. This classifier was later used to detect the graspable regions of novel objects. At the end, a simple lookup-table procedure were used to lift the objects placed in different orientations.

A. Touch region detection results

The robot initially has a rough grasping skill; it reaches for the center of the object and encloses its fingers upon contact. A human caregiver interferes with the execution of this basic skill in attempt to achieve successful grasping. The human provides only partial guidance, making this a true collaboration. In this setup, learning of caregiver learning is also critical, as the ability of the robot control system and the properties of the robot hand for grasping must be learned by the caregiver. Once the human-robot collaborative system manages grasping, by using camera and proprioceptive information the robot can discern parts of the object that it grasps. This provides valuable information for the robot to develop its perceptual system for both action planning and recognition. For example, a bottom-up attention system can be transformed into a system where top-down biases can be also considered (e.g. searching for a handle). In the current implementation, we show that this information can be used to develop 'handle detectors' so that the robot can perform general handle-grasps without human guidance. Fig. 3 shows the percept of the robot before action execution (the top panel). The successful grasping obtained by human guidance is detected by the robot and the positions of the fingers on the object are computed as illustrated in Fig. 3. Repeated application of this process, allows robust discrimination of object points that afford the current action (i.e. handle grasping).

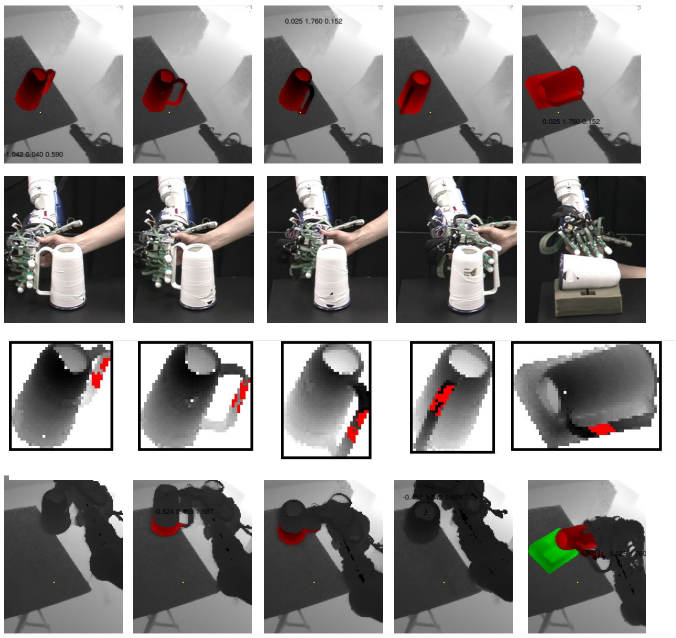

Fig. 3. Guided grasp experience. The figures show the range images measured by the range camera where the depth of the points are encoded in grayscale. The top row shows the snapshots of 5 different environments where the detected objects are shown in red. The second row shows how robot hand was controlled to grasp the objects. The third row focuses on the object and shows the 'first-contact' points. The bottom row gives the snapshots after each behavior execution. The robot can understand the success of grasping by checking the table surface and the force sensor readings on the wrist. Movie is available at http://emreugur.net/movies/humanoids2011

B. Grasp region detection on novel objects

We tested this classifier using grasp executions mediated by the caregiver for a mug type object that is placed at five different orientation as shown in Figure 4 (a). The histograms for each of these executions are given in Figure 4 (b) and (c) corresponding to handle and body histograms, respectively To the human eye it is evident that the distributions are different. The final representative distributions for handle (p_H) and body (p_B) obtained by combining these individual histograms as indicated in Equation 2 are shown in Figure 5.

We first, tested whether this simple classifier can identify the handle parts of the original object accurately or not. In particular, during an interaction not all parts of the handle are

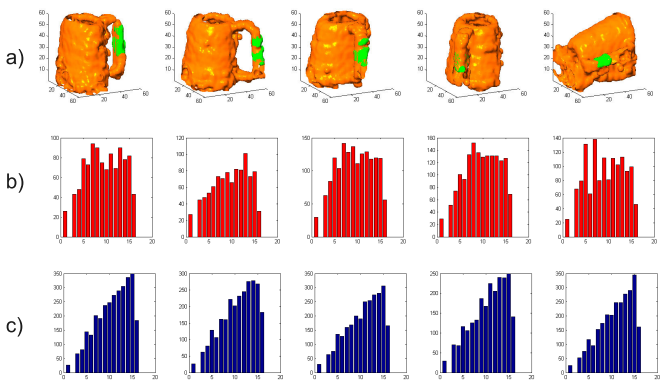

Fig. 4. 'First-touch' regions. (a) shows the objects and their first contact voxels obtained during training (see Figure 3). (b) and (c) gives the corresponding averaged distance histograms for graspable and remaining voxels, respectively. For (a) we generated a surface mesh to cover the voxels and apply smoothing for better presentation of the results.

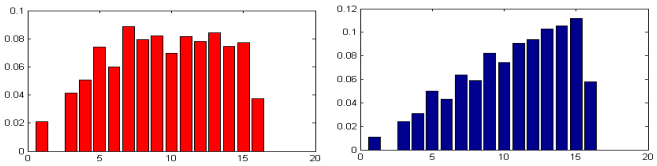

Fig. 5. The mean histograms of graspable and remaining voxel distances, respectively.

touched so they were initially marked as belonging to object body (Figure 4 (a)). With this density based classifier it can be seen that most of the handle voxels are indeed found as handle voxels (Figure 6 (a)) with little false positives. The more challenging task was to see whether this classifier could detect handle-like parts from unseen objects. For this, we used five different objects as seen in Figure 6 (b). Although there were false matches most of the voxels identified as handles were indeed handles or could be considered handles (Figure 6 (c)).

C. Autonomous grasp executions

The focus of this paper was to learn and infer the graspable parts of the objects in our parental scaffolding framework. Although it is not the main focus, the learned knowledge can also be utilized to autonomously control robot arm and to lift the objects. For this purpose, we designed a simple lookup-table based mechanism to select a *reach-grasp-lift* execution trajectory to lift the object that had been used in training.

During training, the robot's guided lift experience was stored as a list of set of object voxels, set of touch voxels, and modified hand-arm angle trajectory. From this experience, the position of the largest touch region relative to the object center was computed. Then, a lookup table was constructed with relative position information in one column and hand-arm trajectory in the other column. When a new object is perceived, the robot first finds the grasp regions using the simple classification method, then computes the position of the largest grasp region relative to the object center. This relative position is searched in the lookup table, the closest

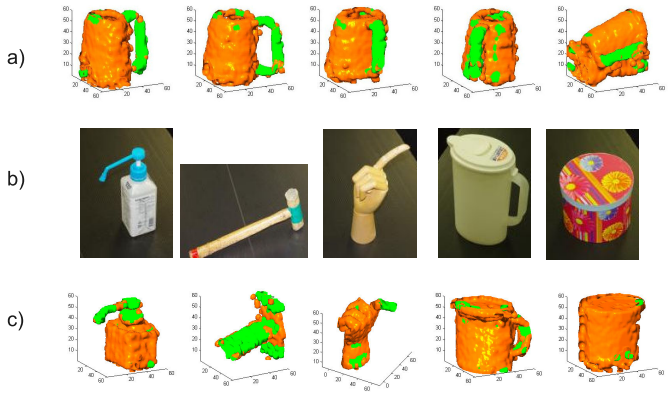

Fig. 6. (a) gives the results obtained from grasp classification of each voxel on training objects. See Fig. 4 for real touch regions of these objects. On the other hand, (c) gives the grasp classification results for novel objects whose pictures were shown in (b). For (a) and (c), we generated a surface mesh to cover the voxels and apply smoothing for better presentation of the results.

experienced relative grasp region position is found, and the corresponding hand-arm trajectory is executed. Note that the use of this distance metric for trajectory selection is limited to the object that was experienced during training, i.e. this metric cannot handle multiple objects with different shapes and sizes.

In the experiments, the object used in training was placed in 5 different orientations. Each row in Figure 7 corresponds to a grasp execution for a different orientation. The snapshots were taken for initial hand posture, while hand was reaching the object, during the first contact, during grasping and at the final stage of lifting, respectively. The first four executions were successful at the end since the object was placed in similar orientations with training instances. In the last execution, the handle was behind the object, so the robot selected an incorrect execution trajectory.

IV. Conclusion

In this paper, we presented a framework that applies the parental scaffolding concept from developmental psychology to a grasping task with a anthropomorphic robot platform. In our framework, robot's default *reach-grasp-lift* trajectory for object center with fixed hand orientation is physically modified online by a human caregiver using the force/moment sensor mounted on the wrist. Utilizing human guidance, the robot was able to successfully grasp the objects, and detect the initial contact points on the objects as graspable regions. Then, it can classify regions of other objects based on the mean statistical distance properties of grasp points on the experienced regions. After a series of guided grasping experiences, the robot was able to (1) detect graspable parts of novel objects and (2) use this information to select grasp trajectories it experienced during (guided) learning and execute them autonomously.

Incorporating offline human judgment in autonomous grasping task has already proven to be a fruitful approach. At one notable study, Saxena et al. [19] trains the robot on synthetic 2D images marking good grasping points. After the training, the robot can identify good grasping points on a novel object,

Fig. 7. Robot grasps objects using the trajectories learned during scaffolding. Each row corresponds to a different grasp execution.
Movie is available at http://emreugur.net/movies/humanoids2011

given only its 2D images. Once the candidate grasping points are identified in 2D, several images are combined to triangulate the points and estimate their 3D positions. [20] represents grasping affordances for a gripper as a continuous distribution in 6D pose space where the distribution is initialized by predefined visual descriptors or sample grips executed by a human (not the robot), and refined through robot's self-exploration.

While these works aim to provide the robot with an initially perfect or almost-perfect information, our focus has been the online intervention of a human teacher to correct the robot's naive movements on-the-fly. A similar approach would have been followed by a mother who watches her infant's play, meddling only when the infant seems in need of help. Although the results demonstrate that the robot was able to efficiently learn of graspable regions through parental scaffolding and was able to detect graspable parts of novel objects, there is room for improvement. First of all, the number of the objects and their variety should be increased. Second, for more realistic and generalized grasp executions, better learning algorithms between grasp region features and grasp behavior execution parameters (such as approach direction or hand posture) should be employed. Finally, from a social sciences perspective, how human caregivers improve in teaching grasp skills to the robot can be studied for improving the human-robot interaction.

ACKNOWLEDGMENTS

This research was supported in part by Global COE Program "Center of Human-Friendly Robotics Based on Cognitive Neuroscience" of the Ministry of Education, Culture, Sports, Science and Technology, Japan. It was also partially funded by the European Commission under the ROSSI project (FP7-21625) and the TÜBİTAK through project 109E033.

REFERENCES

[1] L. E. Berk and A. Winsler, "Scaffolding children's learning: Vygotsky and early childhood education," in *National Assoc. for Education*, 1995.
[2] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem-solving," *Journal of Child Psychology and Psychiatry*, vol. 17, pp. 89–100, 1976.
[3] P. Zukow-Goldring and M. A. Arbib, "Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention," *Neurocomputing*, vol. 70, p. 21812193, 2007.
[4] N. Goubeta, P. Rochat, C. Maire-Leblond, and S. Poss, "Learning from others in 9–18-month-old infants," *Infant and Child Development*, vol. 15, pp. 161–177, 2006.
[5] R. M. Hodapp, E. C. Goldfield, and C. J. Boyatzis, "The use and effectiveness of maternal scaffolding in mother-infant games," *Child Development*, vol. 55, no. 3, pp. 772–781, 1984.
[6] C. Breazeal, "Learning by scaffolding," Ph.D. dissertation, Elec. Eng. Comp. Sci. MIT, Cambridge, MA, 1999.
[7] J. Saunders, C. L. Nehaniv, and K. Dautenhahn, "Teaching robots by moulding behavior and scaffolding the environment," in *Proceedings of the ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, p. 118125.
[8] R. J. Brand, D. A. Baldwin, and L. A. Ashburn, "Evidence for 'motionese': Modifications in mothers' infant-directed action," *Developmental Science*, vol. 5, no. 1, pp. 72–83, 2002.
[9] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 44–54, May 2009.
[10] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *IEEE International Conference on Robotics and Automation*, 2008, pp. 3545–3550.
[11] Y. Nagai and K. J. Rohlfing, "Parental action modification highlighting the goal versus the means," in *IEEE* 7^{th} *International Conference on Development and Learning*, 2008.
[12] B. D. Argall, E. L. Sauser, and A. G. Billard, "Tactile guidance for policy refinement and reuse," in *Proceedings of the 9th IEEE International Conference on Development and Learning*, 2010, pp. 7–12.
[13] A. L. Thomaz and C. Breazeal, "Robot learning via socially guided exploration," in *Proceedings of the 6th International Conference on Development and Learning*, 2007, pp. 82–87.
[14] J. Gibson, *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1986.
[15] E. Ugur, E. Oztop, and E. Sahin, "Goal emulation and planning in perceptual space using learned affordances," *Robotics and Autonomous Systems*, vol. 59, no. 7–8, pp. 580–595, 2011.
[16] E. Ugur, E. Oztop, and E. Şahin, "Going beyond the perception of affordances: Learning how to actualize them through behavioral parameters," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 4768–4773.
[17] A. Murata, V. Gallese, G. Luppino, M. Kaseda, and H. Sakata, "Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip," *Journal of Neurophysiology*, vol. 83, no. 5, pp. 2580–2601, MAY 2000.
[18] H. Sakata, K. I. Tsutsui, and M. Taira, "Toward an understanding of the neural processing for 3d shape perception," *Neuropsychologia*, vol. 43, pp. 151–161, 2005.
[19] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.
[20] R. Detry, E. Başeski, M. Popović, Y. Touati, N. Krüger, O. Kroemer, J. Peters, and J. Piater, "Learning continuous grasp affordances by sensorimotor exploration," in *From Motor Learning to Interaction Learning in Robots*, ser. Studies in Computational Intelligence. Springer Berlin / Heidelberg, 2010, vol. 264, pp. 451–465.