Proceedings of the 2008 IEEE International Conference on Robotics and Automation
pp. 3545-3550, May 2008

Toward Designing a Robot that Learns Actions from Parental Demonstrations

Yukie Nagai, Claudia Muhl, and Katharina J. Rohlfing

***Abstract*— How to teach actions to a robot as well as how a robot learns actions is an important issue to be discussed in designing robot learning systems. Inspired by human parent-infant interaction, we hypothesize that a robot equipped with infant-like abilities can take advantage of parental proper teaching. Parents are known to significantly alter their infant-directed actions versus adult-directed ones, e.g. make more pauses between movements, which is assumed to aid the infants' understanding of the actions. As a first step, we analyzed parental actions using a primal attention model. The model based on visual saliency can detect likely important locations in a scene without employing any knowledge about the actions or the environment. Our statistical analysis revealed that the model was able to extract meaningful structures of the actions, e.g. the initial and final state of the actions and the significant state changes in them, which were highlighted by parental action modifications. We further discuss the issue of designing an infant-like robot that can induce parent-like teaching, and present a human-robot interaction experiment evaluating our robot simulation equipped with the saliency model.**

I. INTRODUCTION

For robots that learn to understand and/or imitate actions from human demonstrators, it is important to be equipped with a learning mechanism appropriately coupled with human teaching. Asada et al. [1] suggested that both the embedded structure of robots and the environmental factors including human teachers should adequately be designed so as to synergistically facilitate robot learning. As seen in parent-infant interaction, a parent (i.e. a teacher) can support his/her infant (i.e. a learner) by adjusting the difficulty of a taught task according to the infant's abilities. The infant, on the other hand, develops his/her perceptual and motion capabilities as he/she grows, which results in the improvement in his/her acquired skills.

On the basis of the idea, Nagai et al. [2] proposed a developmental learning model for robotic joint attention. Their robot able to develop its visual capability learned the task from a human caregiver while the caregiver adaptively evaluated the robot. Specifically, the criterion for evaluating the task was modified as the robot improved the accuracy of the task. Their comparative experiment showed that the model enabled the robot to acquire the joint attention skill faster and in a better manner than models without any adaptability. Uchibe et al. [3] also demonstrated the effectiveness of developmental mechanisms. Their robot, capable of attuning its internal state to the complexity of the environment, could learn to play soccer whereas other robots without such an ability could not. Yoshikawa et al. [4] showed that people's adaptability in teaching enabled their vocal robot to learn to produce vowels. Their robot equipped only with an immature capability (i.e. the ability to produce random sounds) could obtain the meaningful category of sounds by being influenced by the human utterance. These studies have demonstrated the validity of adaptive teaching appropriately coupled with robot learning; however, there is still an open question as to whether such teaching-learning scenarios can be established in natural human-robot interactions (HRI). Since in their experiments the designers taught the tasks to their robots, we should wonder how naive people want to teach a robot and whether a robot can induce their proper teaching. Breazeal and her colleagues [5]–[7] have addressed these questions in reinforcement learning scenarios, and suggested that the social cues presented by a robot (e.g. gaze behaviors and gestures to express the robot's uncertainty about a task) can make the teaching and learning more efficient. Our focus compared to their work is on action learning through demonstration.

Inspired by human parent-infant interaction, we hypothesize that a robot equipped with infant-like abilities can take advantage of parental action teaching. Parents are known to significantly alter their actions when interacting with infants compared to when interacting with adults [8]–[13]. They, for example, put longer and more pauses between actions, exaggerate actions, and decompose a rounded movement into several linear movements, which are assumed to help the infants' understanding of the actions. Infants, on the other hand, have little semantic knowledge about the actions and the environment. It thus makes difficult for them to determine where to attend when observing the demonstrated actions. Our hypothesis is that such an immature attention capability induce parent-like action teaching and the parental teaching helps a robot to detect significant information of the actions. This paper presents our two experiments evaluating parental actions directed to infants and an infant-like attention mechanism in HRI.

The rest of this paper is organized as follows: In Section II, we introduce an attention mechanism simulating the bottom-up attention like the infant's. The model, proposed by Itti et al. [14], [15], is driven only by the saliency derived from the primitive features of an image, but enables a robot to detect likely important locations in the scene. Our experiment analyzing parental actions using the model is described in Section III. Section IV presents our idea on designing an infant-like robot that can induce parental teaching. Our preliminary HRI experiment shows that a robot equipped

Y. Nagai, C. Muhl, and K. J. Rohlfing are with the Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany {yukie, cmuhl, rohlfing}@techfak.uni-bielefeld.de

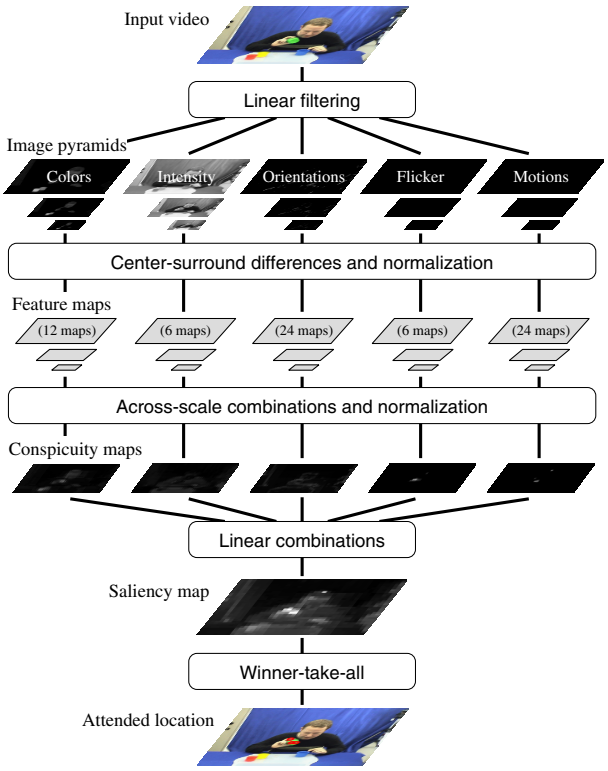


Fig. 1. A model of saliency-based visual attention [14], [15]

with the saliency model can encourage people to properly teach it. Finally, Section V gives the conclusion of our current state and future issues.

II. A MODEL OF SALIENCY-BASED VISUAL ATTENTION

A. Our Assumption

A difficulty in robot action learning is that a robot does not know what visual aspects it should attend to although exposed to a huge amount of sensory information. In contrast to the traditional approaches in which the visual locations and/or the features to look at were defined beforehand, we suppose that a robot has not been provided with any a priori knowledge about the actions, the environment, or even the human demonstrator.

B. Architecture of the Model

To cope with the challenge, we adopt an attention model based on saliency [14], [15] for the robot's vision. Fig. 1 shows the architecture of the model used in our experiment. The model, inspired by the behavior and the neuronal mechanism of primates, can detect salient locations in a scene, which are outstanding from the surroundings with respect to the color, the intensity, the orientation, the flicker (i.e. change in the brightness), and the motion (i.e. optical flow). The former three static features are extracted in two channels (red/green and blue/yellow), in one (black/white), and in four (0/180, 45/225, 90/270, and 135/315 [deg]), respectively. The latter two motion features are extracted in one (on/off) and in four (same as the orientation), respectively.

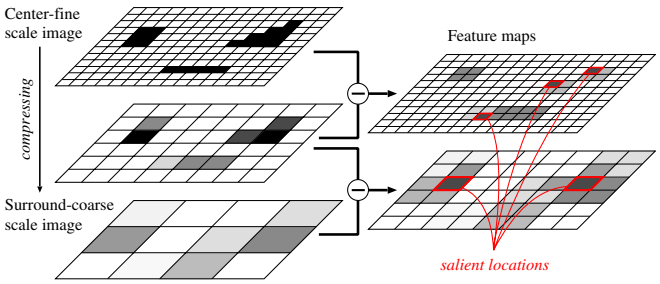


Fig. 2. Calculation for saliency defined as the center-surround difference

An essential processing of the model is the calculation for saliency. Fig. 2 illustrates an example for the intensity feature. The saliency is defined as the difference between the brightness for an image pixel and that for the surroundings, which is calculated by subtracting a center-fine scaled image from a surround-coarser one. In the right of Fig. 2, three locations in the upper feature map and two in the lower map are detected as relatively salient locations. The maps with different scales are then linearly combined across the scales first and next across the features to determine the most salient locations to attend to. Refer to [14], [15] for a more detailed explanation.

C. A Sample Scene from the Experiment

Fig. 3 presents a sample scene from the experiment: (a) shows the attended locations denoted by red circles in the input image (320 $\times$ 256 [pixels]) and (b) shows the corresponding saliency map (40 $\times$ 32 [pixels]). The map was created by linearly combining five conspicuity maps: (c) the color, (d) the intensity, (e) the orientation, (f) the flicker, and (g) the motion map. The brightness for the maps represents the degree of saliency, i.e. white indicates high saliency while black low. In this scene, the color map shows high saliency for the green, the yellow, and the red cup as well as for the person's face and hands. The intensity map presents high saliency for the white tray and the person's black clothes. The orientation map, in contrast, detects the person's face, his hands, and the contour of the tray because of their rich edges. Both the flicker and the motion map show significant saliency for the moving locations, i.e. the person's right hand with the green cup and his face. As the result, three attended locations on the green cup or on the person's right hand were detected from the saliency map, which equally summed up the five conspicuity maps. In our experiment, the locations for which saliency was higher than 0.9 $\times$ the maximum in each image frame were selected as the attended locations.

III. QUANTITATIVE ANALYSIS OF PARENTAL ACTION DEMONSTRATION

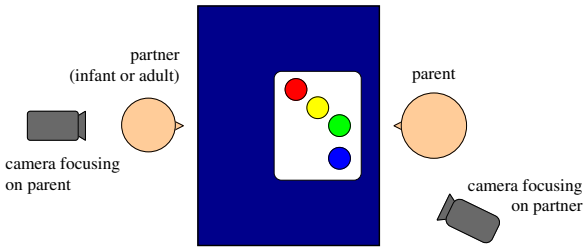
To investigate how parental action demonstrations can help a robot to learn the actions, we analyzed videotaped data of parent-infant/-adult interactions by applying the saliency-based attention model. In contrast to the former study [12],

(a) attended locations indicated in the input image (b) saliency map (sum of (c)-(g))

(c) color map (d) intensity map (e) orientation map

(f) flicker map (g) motion map

Fig. 3. An example of the attended locations and the corresponding saliency map combining equally the five conspicuity maps



(a) top-view of experimental setup

(b) camera image focusing on a parent (c) camera image focusing on an infant

Fig. 4. Experimental setup and sample image frames of videos

which focused only on the task-related actions, our analysis dealt with all visual aspects of the interactions.

A. Subjects and Procedure

Subjects were 15 parents (5 fathers and 10 mothers) of preverbal infants at the age of 8 to 11 months ($M = 10.56,\ SD = 0.89$). Infants of this age were chosen because they are known to be able to understand goal-directed actions [16] and to imitate simple means-end actions [17].

Fig. 4 (a) illustrates the top-view of the experimental setup, and (b) and (c) show sample image frames focusing either on a parent or on an infant. The parents were instructed to demonstrate a stacking-cups task, i.e. sequentially picking up the green, the yellow, and the red cup and putting them into the blue one, to an interaction partner. The partner was first their infant (IDI: Infant-Directed Interaction) and then an adult (ADI: Adult-Directed Interaction). Nothing about the usage of gesture or speech was instructed to the parents, which means that they could interact with the partner as natural as usual.

B. Comparative Analysis between IDI and ADI

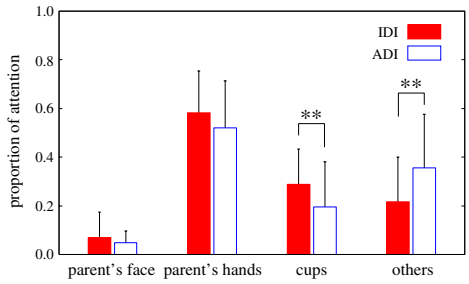
We analyzed the videos recording the parental actions as shown in Fig. 4 (b). The videos were fed into the saliency model, and the locations attended to by the model were evaluated afterward. In order to examine how much important information was detected in IDI compared to in ADI, the attended locations in each condition were categorized into four groups: the parent's face, his/her hands, the cups, or others (e.g. the parent's clothes and the tray). The categorization was automatically done by examining the color and the position in the image (e.g. the parent's hands should mostly be detected lower than his/her face). For example, in Fig. 3 (a) the upper attended location was categorized into the parent's hands whereas the lower two were the cups.

Figs. 5 (a) to (c) show the proportions of the attended locations in the before-, during-, and after-task phase. The filled and open bars present the results for the IDI and the ADI condition. The beginning and end of the during-task phase were defined as when a parent picked up the first cup and when he/she put down the final cup into the blue one, respectively. The each length for the before- and after-task phase was 2 [sec].

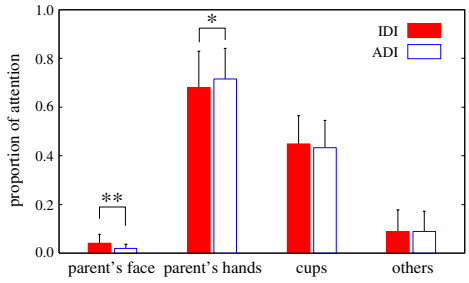
C. Result I: Highlighting the Initial and Final State of the Cups

Our first analysis focusing on the cups revealed that they were attended to more often in IDI than in ADI before the task started and after it fulfilled. A non-parametric test (Wilcoxon test) on the result for the before-task phase showed a significant difference between the IDI and the ADI condition (the third left in Fig. 5 (a); $Z = -2.045,\ p < 0.05$). As for the after-task, a parametric t-test showed a trend between the two conditions (the third left in Fig. 5 (c); $t(14) = 1.846,\ p = 0.086$). These results indicate that the initial and final state of the cups were highlighted by parental action modifications.

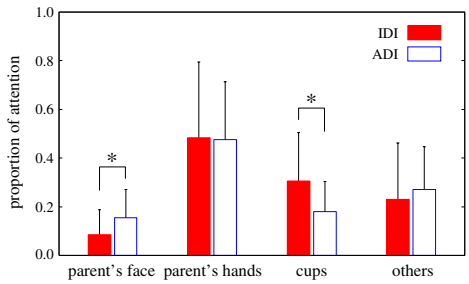
In IDI the high saliency for the cups was caused by two types of parental behaviors: suppressing their body movement or adding movement to the cups. Many parents took a long pause before starting the task as well as after fulfilling it. They completely stopped their movement and closely looked at their infant to examine whether the infant had been being engaged in the interaction. This behavior



(a) before task (2 [sec])



(b) during task



(c) after task (2 [sec])

Fig. 5. Proportions of attended locations (**: significant difference, *: statistical trend)

made the cups relatively salient because they were already conspicuous in terms of the color and the orientation. Some parents, by contrast, generated additional movements with the cups. Specifically, they picked up a cup and shook it before starting the task, which seemed to try to directly draw the infant's attention to it. For the saliency model, this behavior brought high conspicuity for the flicker and the motion channel enough to attract the model's attention. Note that this behavior was not included in the during-task phase because it was irrelevant to achieving the task.

In contrast to IDI, the parents in ADI did not highlight the cups or even other task-relevant locations. They just began demonstrating the task without taking a pause or shaking a cup. The reason is considered that the adult partner was supposed to be easily able to focus on the task-relevant locations. As a supporting result, we found a significant ADI-IDI difference in the attention proportion for the others in the before-task phase (the rightmost in Fig. 5 (a); $Z = -1.988,\ p < 0.05$). This result indicates that in ADI the parents did not make effort of highlighting the task-related locations.

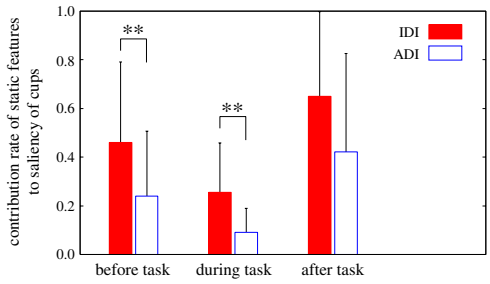


Fig. 6. Contribution rate of static features to saliency of cups

D. Result II: Indicating Significant State Changes in the Task

Our second analysis focusing on the parent's face showed that it attracted the model's attention more often in IDI than in ADI during the task, whereas the contrary was shown after the task. The non-parametric test on the results for the during-task and the after-task phase revealed a significant difference (the leftmost in Fig. 5 (b); $Z = -2.556,\ p < 0.05$) and a statistical trend (the leftmost in Fig. 5 (c); $Z = -1.874,\ p = 0.061$) between IDI and ADI, respectively.

In IDI the parents tended to address the infant while pausing the task demonstration. During executing the task, they sometimes stopped their cup-handling movement, and then commented on the action and/or showed emotional facial expressions to maintain the infant's attention. This behavior caused relatively high saliency for their face enough to attract the model's attention. Here we found two types of parental behaviors: indicating a significant state change in their action beforehand and afterward. In the stacking-cups task, putting a cup into another yields a significant change in the visual state. Some parents alerted the infant to this event by pausing their cup-handling movement just before demonstrating it. Others took a long pause after demonstrating it and then commented on it, seeming to tell the infant the sub-goal of the actions. We thus suggest that the parental actions have the effect of providing social signals indicating the significant state changes in the actions.

In ADI, in contrast, the parents rarely paused their hands' movement during executing the task. They kept demonstrating the task, and some of them presented additional gestures to explain the task. Therefore, although most parents verbally commented on the task through the demonstration, their face was not so salient as to attract the model's attention. Instead, in the during-task phase their hands attracted more attention in ADI than in IDI. The non-parametric test revealed a statistical ADI-IDI trend in the attention proportion for the hands (the second left in Fig. 5 (b); $Z = -1.817,\ p = 0.069$).

E. Result III: Emphasizing the Property of the Cups

Our third analysis focused on the contribution of the static features (i.e. the color, the intensity, and the orientation) to the saliency for the cups. We consider that in action learning, not only the trajectory of the motion but also the means of the action and the property of the target should be attended

to. We thus evaluated how much the parental actions could highlight the property of the cups.

Fig. 6 shows the contribution rates of the static features to the saliency for the cups. Higher rates mean that the color, the intensity, and the orientation feature contributed more to the saliency than the flicker and the motion. The rates were examined only when the cups were attended to by the model. The non-parametric test revealed a significant IDI-ADI difference in the before-task phase (the leftmost in Fig. 6; $Z = -2.040,\ p < 0.05$) and also in the during-task phase (the second left in Fig. 6; $Z = -3.045,\ p < 0.05$). These results indicate that in the two phases the parental actions emphasized the property of the cups.

The reasons are considered as follows: As described in Sections III-C and -D, in IDI the parents tended to pause their body movement before starting the task and even during executing the task. In the before-task phase, they took a long pause to emphasize the initial state of the cups. This behavior made the static features contribute more to the saliency for the cups. Similarly, the parents sometimes stopped their cup-handling movement while demonstrating the task. In order to indicate the significant state changes in the demonstration, they paused their hands' movement and verbally addressed the infant. They also tried to highlight the cups by completely suppressing their body movement. This behavior consequently made the property of the cups more visible than the movement. In ADI, by contrast, the static features did not so much contribute to the saliency for the cups, but rather the dynamic features, i.e. the flicker and the motion, did. The reason is that in ADI the parents did not make so much effort of highlighting the cups by pausing their movement, but kept demonstrating the actions. Therefore, although in the during-task phase the cups attracted the model's attention as much as in IDI (see the third left in Fig. 5 (b)), their property were not visible to the model.

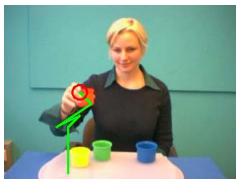
IV. A ROBOT THAT ELICITS PARENT-LIKE ACTION DEMONSTRATION

A. Basic Idea for Designing an Infant-like Robot

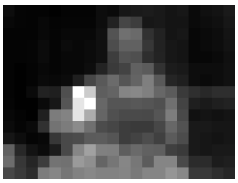
To take advantage of the parental action demonstrations, a robot is desired to be able to motivate human partners to properly teach it as parents do an infant. We propose that not only the infant-like appearance of a robot but also the attentional response based on the saliency can induce parental actions. Infants are supposed to have little semantic knowledge about the actions and the objects used in the actions. Hence, it is difficult for them to predict the following actions and the goal of the actions, which makes them react reflexively. The visual attention of infants is also easily engaged by distractions even when they are interacting with their parents. They seem to rely more on the primal visual information than on the higher knowledge about the actions compared to adults. Schlesinger et al. [18] demonstrated that the saliency model can simulate the infant attention in a perceptual completion task. We hypothesize that the saliency model enables a robot to be recognized as an infant-like agent and consequently encourage human partners to properly teach it actions.

(a) a sample scene of experiment

(b) robot simulation



(c) attended location



(d) saliency map

Fig. 7. HRI experiment using a robot simulation equipped with the saliency-based attention model

B. Preliminary Experiment of HRI

To evaluate our hypothesis, we implemented the saliency model into a robot simulation and conducted an HRI experiment. Fig. 7 (a) shows a sample scene of the experiment, in which a human partner was demonstrating a stacking-cups task to our robot simulation. Fig. 7 (b) shows the robot responding to the partner's action, and (c) and (d) show its attended location and the corresponding saliency map. The robot was displayed on a computer monitor placed in front of the partner, and a camera for the robot's vision was put on it. The robot was programmed to look at the most salient location in a scene although the attention mechanism was not instructed to partners. The partners instead could only recognize that the robot was gazing at something interesting in the environment responding to their actions. 22 naive people (16 university students major in computer science and 6 in humanities) participated in the experiment.

The experiment was originally conducted to evaluate the effects of disturbance in HRI [19]. The robot's attention as well as the infants' can easily be distracted from the interaction by being presented with a more salient object. We thus superimposed a salient target in the robot's vision, and investigated how the distracted robot's attention influenced the partners' actions.

C. Findings by Qualitative Analysis

Our qualitative evaluation using the ethnomethodological conversation analysis revealed that the attentional response of the robot motivated people to properly teach it the actions. Some participants carefully examined the robot's attention and modified their actions so that the robot could easily follow the actions. They, for example, approached to the

robot and closely showed an object as parents do to an infant. They also exaggerated their actions by making pauses, shaking an object, and amplifying their body movement. These phenomena were observed especially after the participants had recognized the distracted attention of the robot. The distraction induced the participants' careful attention and the modifications in their actions so that they could see the robot responding to their actions. Although these findings are still preliminary and have not been quantitatively evaluated yet, we consider that the saliency-based visual attention has the potential to make a robot be accepted as an infant-like agent.

V. CONCLUSION AND FUTURE ISSUES

We presented two experimental results to support our hypothesis that a robot equipped with infant-like abilities can take advantage of parental teaching. In parent-infant interaction, parents aid infant learning by adjusting their teaching strategies while infants' immaturity influences the parents' actions. Our first analysis focusing on parental action demonstration revealed their effect of highlighting the meaningful structures of the actions. The important structures became salient enough to be extracted by a primal attention model. Our second experiment concerning the design issue of an infant-like robot showed a positive effect of the attention model. Our robot simulation equipped with the model induced parent-like action demonstrations of human partners.

These results are an important first step to reach our goal, i.e. designing a robot that learns actions from parental demonstrations. However, there are still open questions concerning the current topics:

- how do parents modify their actions when demonstrating different tasks to infants?
- how do other factors except the robot's visual attention influence the people's actions in HRI?

Regarding the first question, we consider that different aspects should be emphasized depending on the tasks. The stacking-cups task, for example, is a goal-oriented one, and thus the initial and final state of the actions were highlighted by the parental modifications. By contrast, in a motion-oriented task such as dancing, the trajectory of the movement should be emphasized so that a learner can reproduce each movement. Our interest is to find out the difference in parental modifications between actions and to investigate the effect in infant/robot action learning. The second question concerns the design of HRI. Studies on social robots have suggested that the appearance and the behavior of a robot as well as the context influence the people's impression of the robot [20], [21]. In our HRI experiment, the infant-like appearance of our robot also influenced the people's actions. Moreover, the particular situation caused by the robot's distracted attention also facilitated the parent-like action modifications. We thus intend to conduct further HRI experiments to examine the effects of each factor.

REFERENCES

[1] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.

[2] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Advanced Robotics*, vol. 20, no. 10, pp. 1165–1181, 2006.

[3] E. Uchibe, M. Asada, and K. Hosoda, "Environmental complexity control for vision-based learning mobile robot," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 1998, pp. 1865–1870.

[4] Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga, "A constructivist approach to infants' vowel acquisition through mother-infant interaction," *Connection Science*, vol. 15, no. 4, pp. 245–258, 2003.

[5] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo, "Tutelage and collaboration for humaoid robots," *International Journal of Humanoid Robots*, vol. 1, no. 2, pp. 315–348, 2004.

[6] A. Lockerd and C. Breazeal, "Tutelage and socially guided robot learning," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

[7] A. L. Thomaz and C. Breazeal, "Transparency and socially guided machine learning," in *Proceedings of the 5th International Conference on Development and Learning*, 2006.

[8] N. Masataka, "Motherese in a signed language," *Infant Behavior and Development*, vol. 15, pp. 453–460, 1992.

[9] J. M. Iverson, O. Capirci, E. Longobardi, and M. C. Caselli, "Gesturing in mother-child interactions," *Cognitive Development*, vol. 14, pp. 57–75, 1999.

[10] L. J. Gogate, L. E. Bahrick, and J. D. Watson, "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures," *Child Development*, vol. 71, no. 4, pp. 878–894, 2000.

[11] 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.

[12] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" *Advanced Robotics*, vol. 20, no. 10, pp. 1183–1199, 2006.

[13] Y. Nagai and K. J. Rohlfing, "Can motionese tell infants and robots "what to imitate"?" in *Proceedings of the 4th International Symposium on Imitation in Animals and Artifacts*, 2007, pp. 299–306.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[15] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proceedings of the SPIE 48th Annual International Symposium on Optical Science and Technology*, 2003, pp. 64–78.

[16] I. Kiraly, B. Jovanovic, W. Prinz, G. Aschersleben, and G. Gergely, "The early origins of goal attribution in infancy," *Consciousness and Cognition*, vol. 12, pp. 752–769, 2003.

[17] J. A. Sommerville and A. L. Woodward, "Pulling out the intentional structure of action: the relation between action processing and action production in infancy," *Cognition*, vol. 95, pp. 1–30, 2005.

[18] M. Schlesinger, D. Amso, and S. P. Johnson, "Simulating infants' gaze patterns during the development of perceptual completion," in *Proceedings of the 7th International Conference on Epigenetic Robotics*, 2007, pp. 157–164.

[19] C. Muhl and Y. Nagai, "Does disturbance discourage people from communicating with a robot?" in *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication*, 2007, pp. 1137–1142.

[20] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication*, 2003, pp. 55–60.

[21] T. Minato, M. Shimada, S. Itakura, K. Lee, and H. Ishiguro, "Evaluating the human likeness of an android by comparing gaze behaviors elicited by the android and a person," *Advanced Robotics*, vol. 20, no. 10, pp. 1147–1163, 2006.