Computational Analysis of Motionese Toward Scaffolding Robot Action Learning

Yukie Nagai and Katharina J. Rohlfing

***Abstract*—A difficulty in robot action learning is that robots do not know where to attend when observing action demonstration. Inspired by human parent-infant interaction, we suggest that parental action demonstration to infants, called *motionese*, can scaffold robot learning as well as infants'. Since infants' knowledge about the context is limited, which is comparable to robots, parents are supposed to properly guide their attention by emphasizing the important aspects of the action. Our analysis employing a bottom-up attention model revealed that motionese has the effects of highlighting the initial and final states of the action, indicating significant state changes in it, and underlining the properties of objects used in the action. Suppression and addition of parents' body movement and their frequent social signals to infants produced these effects. Our findings are discussed toward designing robots that can take advantage of parental teaching.**

***Index Terms*—Bottom-up visual attention, motionese, parental scaffolding, robot action learning.**

I. Introduction

LEARNING actions from human demonstrators is an important ability for robots that have been designed to interact with humans. As in human society, people's skills are passed on to others through demonstrations [1], robots which are able to learn new tasks by observing them can facilitate human-robot interaction and accelerate the teaching-learning processes. In such a scenario, however, robots encounter a problem of not knowing relevant features of the task. When observing a task demonstration, robots are exposed to a huge amount of sensory information, some of which are important in achieving the task but some of which are not. In order to learn to perform the task, robots have to appropriately select the relevant information while paying no or less attention to irrelevant. This issue is stated as "what to imitate" in the studies of robot imitation, i.e., what aspects of the demonstrated action robots should look at and what they should reproduce in imitation [2]–[6]. Robots which have no *a priori* knowledge about the task, the environment, nor even the human demonstrator must overcome the challenge of detecting the relevant information.

Manuscript received November 18, 2008; revised February 21, 2009. First published April 17, 2009; current version published May 29, 2009. The work of Y. Nagai was supported by Honda Research Institute Europe. The work of K. J. Rohlfing was made possible by the Dilthey Fellowship (Volkswagen Foundation).

The authors are with the Research Institute for Cognition and Robotics, Bielefeld University, 33594 Bielefeld, Germany (e-mail: yukie@techfak.uni-bielefeld.de; rohlfing@techfak.uni-bielefeld.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2009.2021090

A promising approach to this issue is to investigate human parent–infant interaction. Infants have little semantic knowledge about the context, the environment, and even the interaction partner. They may not know what their parents try to teach to them and what is demonstrated to them. Therefore, it is difficult for infants as well as for robots to extract the relevant features from their perceptual signals. Nevertheless infants acquire many skills through interactions with their parents. A key factor, which makes it possible, is scaffolding provided by parents. It is known that in action demonstration to infants, parents significantly modify their body movement compared to adult-directed demonstrations (e.g., [7] and [8]). They, for example, use a wider range of movement, repeat the same movement, and make longer and more pauses between movements. This action modification, called *motionese*, is assumed to maintain the infants' attention and support their processing of the action, which leads to their better understanding of the action (see Section II-B for a more detailed review of motionese).

Inspired by the parent–infant interaction, we suggest that motionese can scaffold robot action learning. Parental action to infants is hypothesized to physically highlight the important aspects of the action and thus properly guide robots' attention. Parental modification such as suppression and addition of their body movement might make the relevant features more salient than irrelevant. For example, exaggerating the task-relevant movement would make it distinguishable from other gestures. Taking a pause between movements might emphasize a significant state in the task (e.g., the initial and final states). Our hypothesis is that such important aspects highlighted by parental actions can be detected by bottom-up visual attention. Even if robots do not know what are relevant or circumstantial, parental modifications would draw the robots' attention to the important aspects of the action. In order to verify our hypothesis, we analyzed parental demonstration employing a computational attention model based on saliency. The model [9], [10] is able to detect outstanding locations in a scene in terms of primitive features, e.g., color, orientation, and motion. That is, the model does not require any *a priori* knowledge about the context nor the environment, but enables robots to detect likely important locations. Our analysis using the saliency model can thus address the question as to how motionese assists robots in learning what to imitate.

The rest of this paper is organized as follows: Section II gives the overview of the related work. It first describes the issue of what to imitate in robot learning and then provides psychological evidence about parental teaching for infants. Section III introduces an attention model based on saliency used in our experiment. It is explained that the model refers only to stimuli driven

information but has the ability to extract likely important information. We also emphasize the potential of the model to simulate the primal attention of infants. Sections IV and V present our analytical experiment on parental action. In order to uncover the effects of motionese, we compared parental demonstrations directed to infants versus to adults. Statistical results followed by our qualitative explanations are presented. Section VI discusses the results toward scaffolding robot action learning. We state how motionese can contribute to robot learning for what to imitate and give our idea to design such an architecture. Finally, Section VII concludes the paper.

II. Related Work and Our Novel Hypothesis

A. "What to Imitate" in Robot Action Learning

What to imitate is an important issue to be discussed in robot imitation as well as in robot action learning [2]–[5]. Robots which have *no a priori* knowledge about the context, the environment, nor even the demonstrator (e.g., a human body model nor skin color information) have to overcome the difficulty in detecting the person who is presenting a task, his body parts engaged in the task, objects used in it, and so on. The decision on whether to reproduce the same body movement as the demonstrator's one, i.e., the means of the task, or to achieve only the goal of it by adopting other movement, is also a challenge to be addressed.

To our knowledge, there are only a few studies tackling this issue. Billard and colleagues [11], [12] adapted a probabilistic method to extract the relative importance of movements. Their method is to first measure all candidate variables (e.g., joint angles and the position of an end-effector) over several demonstrations and then extract the key features as the variables with small variances across the demonstrations. Their experiment showed that their robot could appropriately reproduce different tasks, which had different constraints, by putting stronger weights on the key movement. Insights from studies on social learning further motivated them to extend their framework so that it could place a human teacher in the robot's learning loop [13], [14]. In their framework, the teacher was allowed to refine the robot's movement through multimodal interaction, e.g., kinesthetic teaching, gazing, and vocalizing, while he or the robot was performing the task. Alissandrakis *et al.* [15] discussed what to imitate as an issue of determining the granularity of movement. They described that in a chess world, the movement of a chess piece can be imitated by another type of piece either at the path level, the trajectory level, or the end-point level. Taken together, these works showed that robots can appropriately select the features to imitate depending on the task constraints *only if* the candidate variables are given to the robots. It thus leaves an open question as to how robots can know such candidates from their huge amount of sensory signals.

We tackle this issue as a robots' attention problem. We assume that *no* knowledge about the task nor the context is given to robots, and thus robots do not know what to attend to when observing action demonstration. Scassellati [16] pointed out that robots can exploit not only the inherent saliency of signals but also social cues given by a demonstrator, the robots' embodiment, and developmental progress both in the robots and the environment. Our approach addresses the question concerning the last factor, i.e., how developmental constraints can overcome the problem. We adopt a bottom-up attention model as a constraint for robots' vision. It allows robots to detect salient locations although there is no guarantee that the locations are relevant. Our key idea is that combining this simple model with parental action demonstration, which provides an environmental constraint, enables robots to learn what to imitate. The next section describes the evidence about parental teaching and gives our hypothesis on scaffolding robot learning.

B. Evidence About Motionese in Parent-Infant Interaction

Developmental studies have revealed that human parents significantly alter their infant-directed actions compared to adult-directed actions. Brand *et al.* [7], who first used the term of motionese, examined how differently parents demonstrated the usage of novel objects to infants compared to adults. They videotaped parents interacting either with an infant or an adult partner and then asked naive experimenters to manually code it. Their statistical analysis showed higher rates of infant-directed action in six dimensions: the proximity to the partner, the interactiveness, the enthusiasm, the range of the motion, the repetitiveness, and the simplification. Their further analysis revealed the adaptation in parental action depending on the age of infants [17] and also the infants' preference of motionese to adult-directed action [18]. Masataka [19] uncovered a similar phenomenon in signed language. He compared the signs presented by deaf mothers to their deaf infant with those to their deaf adult friend. Similarly to the findings in action demonstration, parental signs were slower, repeated more often, and exaggerated in infant-directed interaction than in adult-directed. His following studies showed that such parental signs attracted greater attention of both deaf and hearing infants [20], [21].

Zukow-Goldring and colleagues [22], [23] investigated how parents guide infants' attention when teaching actions in natural interaction. They focused on the fact that parents do not only demonstrate the movement to infants but also invite them to perform the task, point to important aspects of the task, and moreover embody the infants, i.e., put the infants' body through the movement. From these observations, they suggested that the parental assists can *educate* infants' attention. This suggestion was enhanced and extended by Yu *et al.* [24]. They directly investigated infants' view by placing a small camera on the infants' forehead. Their embodied approach demonstrated that parental actions reduced the uncertainty in the world and moreover the infants themselves did. The limited view of the infants as well as their action to handle an object decreased the ambiguity of the visual input, which experimentally supported Scassellati's discussion [16] on developmental constraints.

Rohlfing and colleagues [8], [25] introduced a computational technique to the analysis of motionese. They employed a 3-D body tracking system, which was originally developed for human–robot interaction [26], to quantitatively analyze parental actions. In their experiment, parents demonstrated a stacking-cups task first to their infant and then an adult partner. Their comparative analysis focusing on the parents'

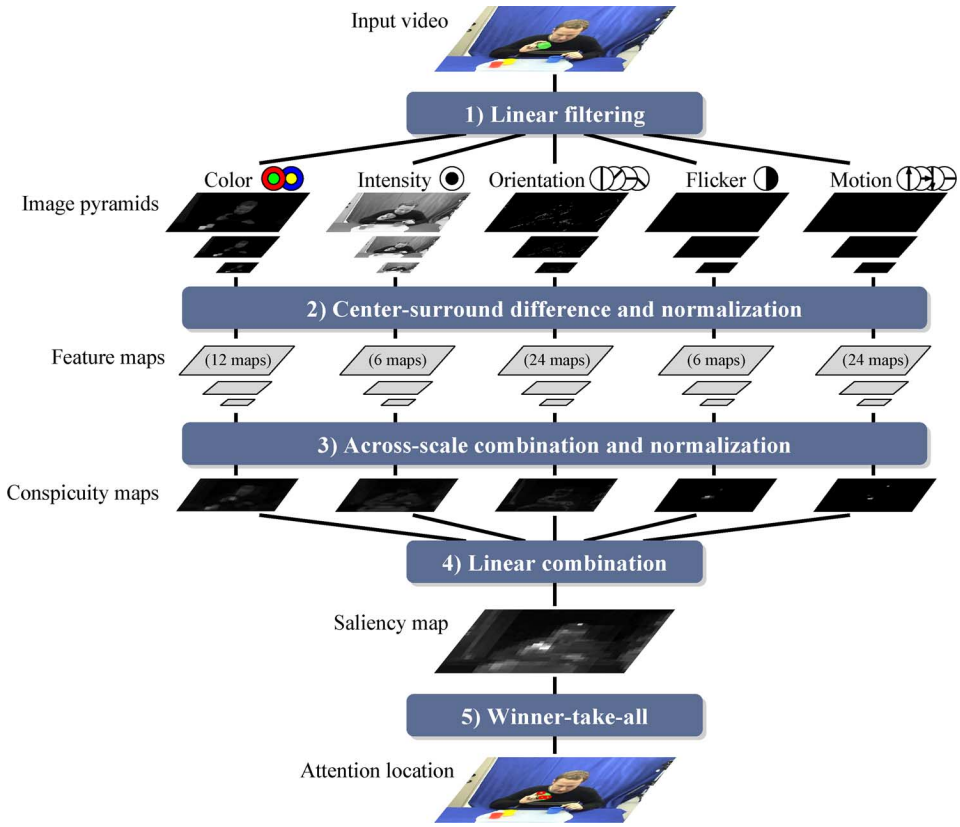

Fig. 1. A bottom-up visual attention model based on saliency.

cup-handling movement revealed that the parents tended to make longer and more pauses between movements and to decompose a rounded movement into several linear movements for the infants. Their suggestion derived from these results is that motionese can support infants and also robots structuring the task-relevant action.

Following the work done by Rohlfing *et al.* [8], [25], we investigate how motionese properly guides robots' attention as well as infants'. In the former study, only the cup-handling movement was analyzed, that is, robots were supposed to know what was relevant to the task. In contrast, we assume that robots do not know what task is demonstrated nor who is demonstrating, and therefore they can rely only on bottom-up signals to control their attention. Our hypothesis is that parental action modifications such as suppression and addition of their body movement change the relative saliency of bottom-up signals so as to draw the robots' attention to the relevant. For example, longer pauses in parents' movement would emphasize the static state of objects, which can impart the goal of the action. The next section introduces a bottom-up attention model used in our experiment.

III. A Bottom-Up Visual Attention Model Based on Saliency

A. Architecture of Model

In order to address the issue of what to imitate, we adopt a visual attention model based on saliency. The model, proposed by Itti *et al.* [9], [10], is inspired by the behavioral and neuronal mechanism of primates and is able to detect salient locations in a scene. The saliency is defined as the outstandingness from the surroundings in terms of primitive features. For instance, a red circle on green background is detected as salient because of its distinctive color. A dot moving left among a number of dots moving right is also salient with respect to the motion direction. In human-robot interaction, the model also enables robots to look at human partners without employing any human model. Because the inherent features of the human face and hands, e.g., skin color and the contours of the eyes, mouth, and fingers, as well as their movement are distinguishable against natural environments, the model can extract them as salient objects.

Fig. 1 presents the architecture of the saliency model. Here we explain the overview of the model. Refer to the original papers [9], [10] for a more detailed.

1) The model first extracts five primitive features: color, intensity, orientation, flicker, and motion, by linearly filtering an input video. The color feature is represented in two channels: red/green and blue/yellow, while the intensity in one: black/white. The flicker channel, which extracts the temporal change in the intensity (e.g., change in the lighting condition), is represented in one: on/off. The orientation (i.e., edges) and the motion features (i.e., optical flow) are represented each in four: 0, 45, 90, and 135 [deg] for the orientation, and 0, 90, 180, and 270 [deg] for the motion. These multiple channels are necessary to discriminate different inclinations of edge features and different directions of movement. Then, all the extracted features create image

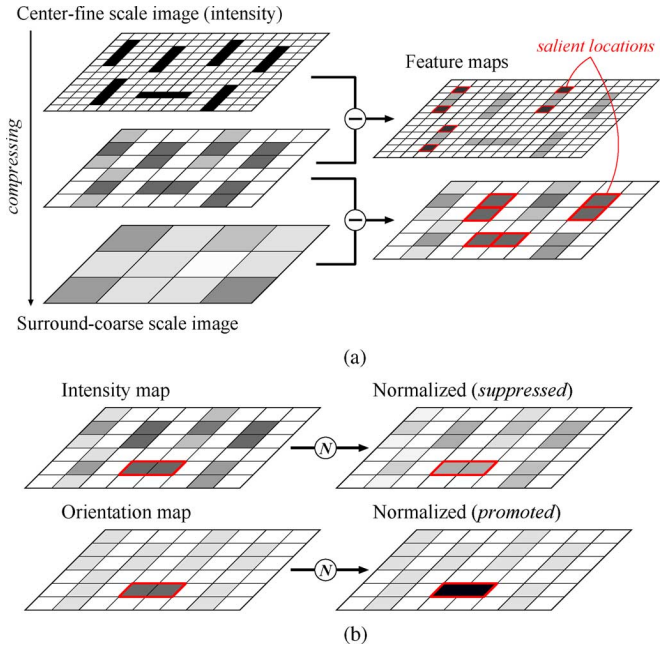

Fig. 2. Two important processes in saliency model. (a) Calculation for center-surround difference. (b) Normalization within feature.

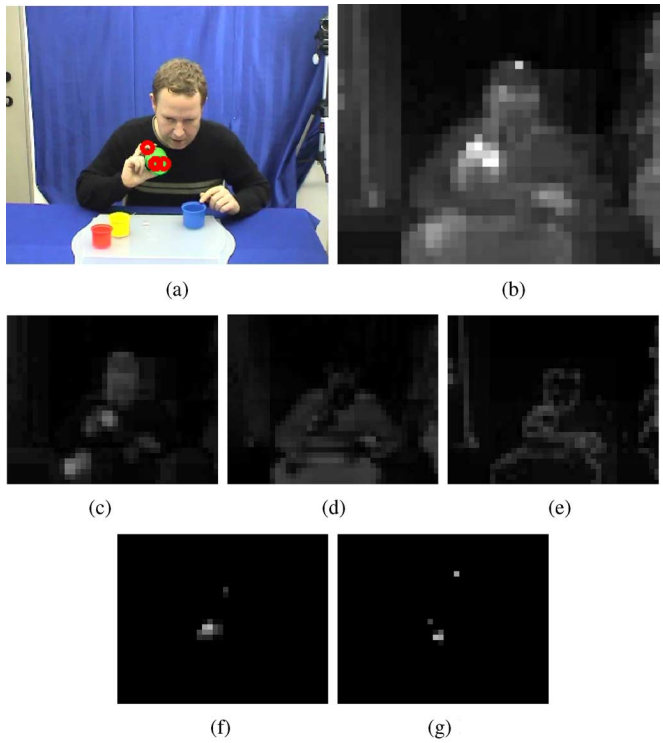

Fig. 3. Saliency maps and attention locations of saliency model. (a) Attention locations of saliency. (b) Saliency map (sum of (c) to (g)) model noted by red circles. (c) Color map. (d) Intensity map. (e) Orientation map. (f) Flicker map. (g) Motion map.

pyramids by compressing the original size of the feature image.

2) For each image pyramid, the difference between a fine-scale image and a coarser-scale image is calculated to obtain the center-surround difference. Fig. 2(a) illustrates the process, where several bars (six vertical bars and one horizontal) are extracted in the intensity image (the top left). The feature maps are created by subtracting a coarser image from a finer one at multiple scales, which allows to detect different sizes of salient objects. In the maps, the bigger the difference is, the more salient the location is.
   The feature maps are then normalized in two steps. First, it eliminates the modality dependence caused by the different amplitude. Secondly, it globally promotes maps containing a few salient locations while suppressing maps with numerous peaks. Fig. 2(b) illustrates the effect, where the intensity and the orientation maps derived from the sample image in (a) are normalized. The intensity map, containing several salient points, is suppressed whereas the orientation map with one salient peak is promoted. As a result, the saliency corresponding to the horizontal bar is enhanced compared to that for the vertical bars.
3) The normalized feature maps are combined across scales and normalized again as in 2). This process generates the conspicuity maps, representing the saliency derived from each image feature.
4) All the conspicuity maps are linearly combined into the saliency map. The obtained map consequently presents the saliency, which takes into account all the features and different sizes of stimuli. Note that when being combined, all the conspicuity maps are weighted equally because adjusting the weights requires context knowledge.
5) Finally, the model selects the locations to attend to based on the saliency map. In our experiment, locations for which the saliency is higher than $0.9 \times$ the maximum in the current frame are selected. That is, it does not simulate gaze shift (i.e., a series of fixations from one location to another) but allows us to find out more general characteristics of input signals from the multiple selected locations.

The model is the same as proposed in [9], [10] except not using the inhibition-of-return mechanism, which simulates an habituation effect by inhibiting the saliency for the attended locations. We decided not to use the mechanism because in parent–infant/human–robot interaction the salient points are always moving in the scene and thus cannot be inhibited by the location.

B. Sample Scene From Experiment

Fig. 3 gives a sample scene from the experiment, where a father is demonstrating a stacking-cups task to his infant. The task involves four colored cups: a red, a yellow, a green, and a blue cup (from left to right in the image), which are placed on a white tray. In the scene, the father is picking up the green cup and presenting it to his infant. Fig. 3(a) shows the locations attended to by the saliency model, which are denoted by red circles in the input image (320×256 [pixel]), (b) the corresponding saliency map (40×32 [pixel]), and (c) to (g) the conspicuity maps with respect to the five features: color, intensity, orientation, flicker, and motion. The saliency map as well as the conspicuity maps denotes the degree of saliency by the brightness of the image, i.e., the higher the brightness is, the more salient the location is.

The resolution of the saliency map was defined following [9], [10], in which the better performance was verified.

In the scene, the green cup as well as the father's right hand holding it is attended to by the model. First, the color channel [see Fig. 3(c)] extracts higher saliency for the green, the yellow, and the red cups as well as for the father's face and hands due to their distinctive colors against the blue background. The blue cup, in contrast, is not so salient in terms of the color because of the similarity to the background. Next, the intensity channel [see Fig. 3(d)] detects the white tray and the father's black clothes as salient features. Because they have extreme brightness or darkness, those features are distinguishable from others. The orientation channel [see Fig. 3(e)] then extracts the father's face and his hands as well as the contours of the cups and the tray as salient locations. Because the face, for instance, contains rich edge features on the eyes, eyebrows, nose, and mouth, those features contribute to the high saliency for it. His hands are also salient due to the edge features detected from the fingers, even when they are not moving. Compared to the above three static features, both the flicker and the motion channels [see Fig. 3(f) and (g)] detect moving locations, i.e., the father's right hand with the green cup as well as his head. Note that the two channels present relatively higher saliency than the static ones because the normalization process promotes the two maps containing a few salient peaks while suppressing the others with numerous peaks. Finally, the saliency map [see Fig. 3(b)], which equally combines the five conspicuity maps, extracts three outstanding locations to attend to: two on the green cup and one on the father's right hand, for which the saliency is higher than $0.9\times$ the maximum in the frame.

C. Capability of Simulating Infants' Attention

Applying the saliency model to the analysis of motionese enables us not only to address the issue of what to imitate but also to uncover infants' learning from parental demonstration. Our hypothesis underlying here is that the saliency model can simulate the primal attention of infants. It is supposed that infants have little semantic knowledge about the context and the environment. In action learning scenarios, infants may not know what parents are demonstrating to them, what the goal of the action is, or what objects are involved in it. Therefore, when deciding where to attend, infants face the same problem as robots do, and have to rely much more on the bottom-up signals than on the top-down information. Golinkoff and Hirsh-Pasek [27] suggested that before 10 months of age, infants are mostly driven by perceptual cues. In word learning, infants rather disregard social cues and use bottom-up salience of an object to associate a word with the object. Schlesinger *et al.* [28] experimentally supported our hypothesis. They demonstrated that the saliency model can reproduce the infants' attention in a perceptual completion task. Our interest, by contract with [28], is on the potential of the model in the context of action learning.

IV. Statistical Analysis of Motionese Employing Saliency Model

We analyzed parental action demonstration employing the saliency model. In order to investigate how motionese can contribute to highlighting the relevant features of the action, we compared parental action in Infant-Directed Interaction (IDI condition) versus in Adult-Directed Interaction (ADI condition).

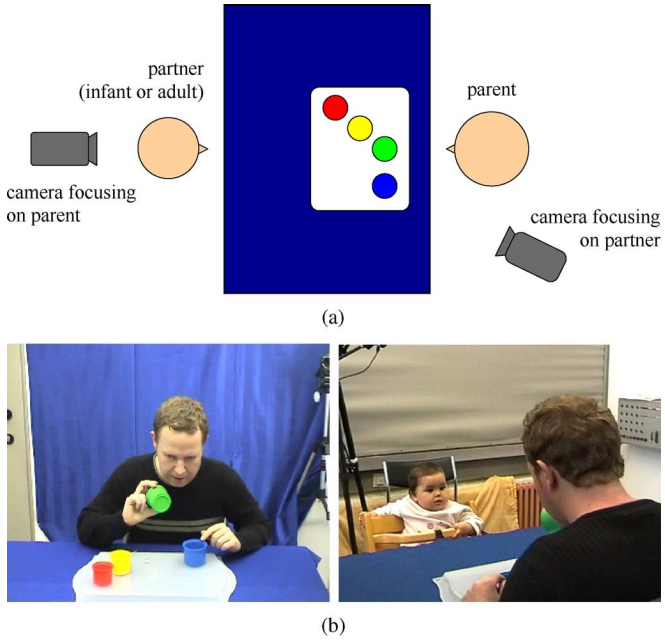

Fig. 4. Experimental setup and sample scene of videotaped interaction. (a) Top-view of experimental setup. (b) Video recording patient's action. (c) Video recording partner's reaction that was used as input to saliency model.

A. Subjects

Subjects were 15 parents (5 fathers and 10 mothers) of preverbal infants. Their infants were 8 to 11 month old ($M = 10.56$, $SD = 0.89$) when they joined the experiment. We chose this age because infants at six months start showing the ability to imitate simple means-end actions [29] and to understand goal-directed actions [30].

B. Experimental Setting and Task

Fig. 4(a) illustrates the experimental setup, and (b) and (c) show the sample images of the videotaped interaction. A parent was seated across a table from an interaction partner. The partner was first his/her infant (IDI) and then his/her spouse (ADI). The parents were asked to demonstrate a stacking-cups task to the partner while explaining how to achieve it. The stacking-cups task was to pick up the green, the yellow, and the red cups and put them into the blue one sequentially. *No* other instruction, for example, on the usage of gestures or speech was given, meaning that the parents could interact with the partners as much as usual. The interaction was recorded by two cameras, one focusing on the parents' action [see Fig. 4(b)] and the other focusing on the partner [see Fig. 4(c)].

C. Analysis of Parental Action Using Saliency Model

We analyzed the videos recording the parents' action as shown in Fig. 4(b). The videos were fed into the saliency model, and the locations attended to by the model were examined afterward. Fig. 3 shows a sample scene from the analysis.

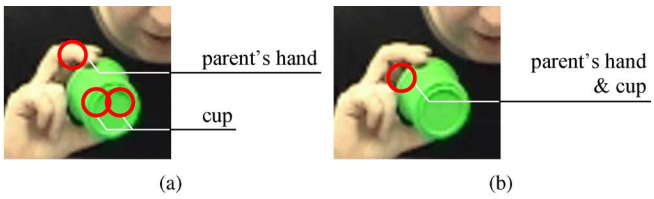

Fig. 5. Classification of attended locations. (a) Attention categorized either as parent's hands or cups. (b) Attention categorized as both parent's hands or cups.

As already explained in Section III-B, the green cup and the father's right hand attracted the model's attention at that moment because of their distinctive color and the edge features as well as the movement.

In order to evaluate how important aspects of the demonstrated action were detected by the saliency model, we classified the attended locations into four regions: the parent's face, his/her hands, the cups, and others (e.g., the parent's clothes and the tray). Fig. 5 gives examples: In (a), the attention locations were categorized either as the parent's hand (the upper one) or as the cups (the lower two), whereas in (b) the location was categorized as the both, i.e., the hands and the cups. Since the saliency was calculated for 8×8 [pixel] of the image (320×256 [pixel] of the input image and 40×32 [pixel] of the saliency map), attention locations sometimes contained two or more features. In such cases, the attention points were allowed to be classified as all the included features. The classification was automatically done by examining the color and the position of the attention. That is, locations which had red, yellow, green, or blue color were categorized as the cups, whereas those with skin color was categorized as the face or the hands. The face and hands were then distinguished by the relative position, i.e., the face should mostly be above the hands.

V. Experimental Results

We compared when and how often the four features, i.e., the parent's face, his/her hands, the cups, and others, attracted the saliency model's attention in IDI versus in ADI. Fig. 6 presents the proportion of the attention in three task-demonstration phases: a) for 2 s before the parents started demonstrating the task, b) while they were demonstrating it (the mean duration was 13 s), and c) for 2 s after they fulfilled it. The duration for the before-/after-task phases was defined by the mean interval between the tasks. The beginning and end of the task were defined as when the parents picked up the first cup and when they put down the final cup into the blue one, respectively. Note that parental actions which involved the cups but irrelevant to achieving the task, e.g., tapping the first cup on the tray to attract the partner's attention, were included in the before-task phase, but not in the during-task phase. The filled bars and the open ones represent the mean proportion of the attention in IDI and in ADI, respectively. The standard deviations are denoted by the error bars. We performed a paired t-test (two-tailed) on the two conditions and revealed significant differences and statistical trends denoted by "++" and "+." Our three main findings are described below.

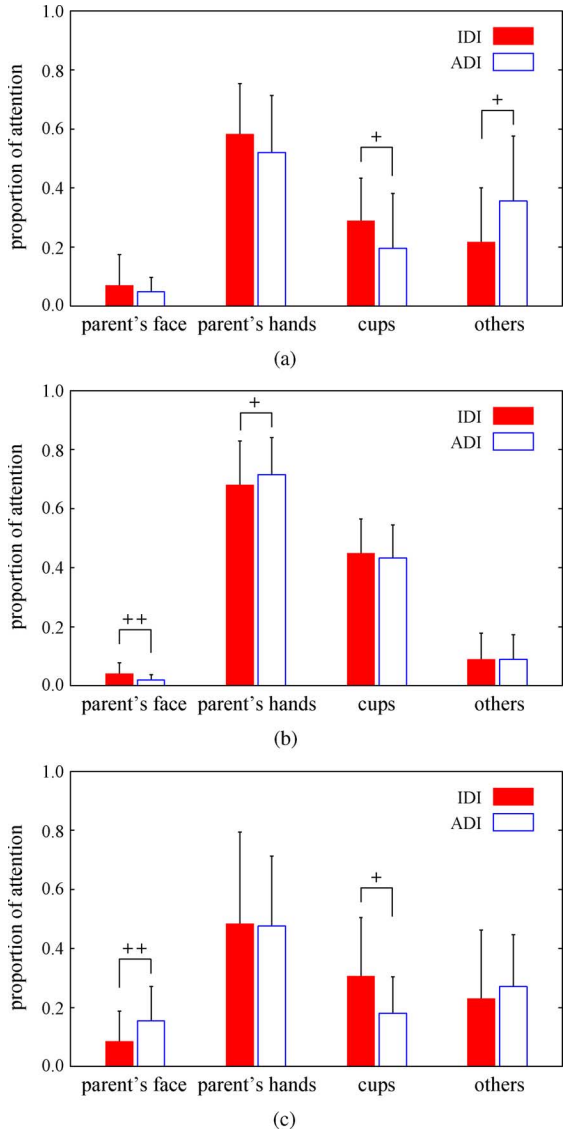

Fig. 6. Proportion of attention of saliency model (++: significant difference, +: statistical trend). (a) Before task (2 [seconds]). (b) During task. (c) After task (2 [seconds]).

A. Highlighting Initial and Final States of Cups

Our first analysis focusing on the cups revealed that they attracted more attention of the saliency model in IDI than in ADI before the task started and after it finished. The paired t-test on the result for the before-task phase revealed a statistical trend between IDI and ADI (the third left in Fig. 6(a); $t(14) = 1.81$, $p = 0.09$). It also showed a statistical trend for the after-task phase (the third left in Fig. 6(c); $t(14) = 1.84, p = 0.08$). These results indicate that the parents tended to highlight the initial and final states of the cups by modifying their action.

In IDI, the high saliency for the cups was caused by two types of parental behaviors: suppressing their body movement and adding movement to the cups. First, we found that many parents took a long pause before starting the task as well as after fulfilling it. They completely suppressed their body movement and closely looked at their infants to examine whether the infants were engaged in the interaction. Fig. 7(a) shows the scenes

(a) (b)

Fig. 7. Parental action highlighting initial and final states of cups in IDI. (a) Taking long pause before (left) and after (right) task. (b) Tapping cup on table before task.

captured when the parents took a pause just before (the left picture) and just after (the right) the task. In both scenes, the cups were attended to by the saliency model. The reason for it is that the suppression of the parents' body movement enhanced the saliency derived from the static features, i.e., color, intensity, and orientation. As seen in the conspicuity maps shown in Fig. 3, the cups were salient with respect to the color and the orientation even without movement. Therefore, when the parents stopped acting, the cups attracted the model's attention due to the inherent saliency. As the second parental behavior, we found that some parents rather generated additional movement to the cups. More specifically, they took the first cup and tapped it on the tray just before starting the task [see Fig. 7(b)]. They seemed to try to draw the infants' attention to the cup by moving it. For the saliency model, this behavior brought the additional saliency derived from the flicker and the motion, which was strong enough to attract the model's attention. Note that even when moving the cup, the parents still kept the position of it so as to teach the infants where the cups were placed, i.e., the initial state of the cups.

In contrast to in IDI, where the parents made much effort to guide the infants' attention, in ADI they rarely did it even before starting the task. They did not make a significant pause in their action nor add movement to the cups. It is assumed that the adult partners could easily recognize the context and thus attend to the task-relevant locations by themselves. Therefore, in ADI the parents might not need to emphasize the initial or final states of the cups. To support this, we found a statistical ADI-IDI trend in the attention proportion for others in the before-task phase (the rightmost in Fig. 6(a); $t(14) = -1.84$, $p = 0.08$), indicating that in ADI, the parents did not make efforts to highlight the task-relevant locations.

B. Frequent Social Signals Indicating Significant Events

Our second analysis focusing on the parent's face revealed higher attention to their face in IDI than in ADI during the task demonstration, whereas the contrary was found after the task. The paired t-test on the result for the during-task phase revealed a significant IDI-ADI difference (the leftmost in Fig. 6(b); $t(14) = 2.63$, $p < 0.05$) and for the after-task phase a significant ADI-IDI difference (the leftmost in Fig. 6(c); $t(14) = -2.10$, $p < 0.05$). This result indicates that the parents gave more frequent social signals to the infants than to the adults.

In IDI, the parents tended to often address the infants during the task demonstration. While performing the task, they sometimes stopped their cup-handling movement, and then commented on it and/or showed emotional facial expressions to maintain the infants' attention. This behavior caused relatively high saliency for their face, enough to attract the model's attention. Here we found two types of parental social behaviors: indicating significant state changes in the task beforehand and afterward. Fig. 8(a) shows the scenes captured when the mothers were verbally addressing their infants shortly before putting down the holding cup. In the stacking-cups task, putting a cup into another yields a significant change in the visual state, i.e., a cup is going to be invisible to the infants. Both of the mothers seemed to alert their infants to the following event by pausing their cup-handling movement and talking to the infants. Fig. 8(b) plots the transition of the model's attention over the task, corresponding to the right mother shown in Fig. 8(a). The three grayed windows and following darker ones denote the period when she was moving a cup from the tray and that when putting it down into the blue one, respectively. That is, the darker windows are corresponding to the significant events. From Fig. 8(b), we can see that her face attracted the model's attention while she was moving a cup, i.e., shortly before each important event. In contrast, some other parents addressed their infants after achieving each event. Fig. 9(a) shows a scene, where the mother was talking to her infant just after placing the red cup into the yellow one. Fig. 9(b) is the corresponding attention transition. More attention to her face after the darker grayed windows indicates that she gave social feedback to her infant after performing each important event.

Although in ADI the parents were also verbally addressing the adult partners through the demonstration, they rarely paused their hands' movement. They kept performing the task, and some of them even produced additional gestures to explain the task. Therefore, their face was not so salient as to attract the model's attention. Instead, their hands attracted more attention in ADI than in IDI during the task. The paired t-test revealed a statistical trend in the attention proportion for the hands (the second left in Fig. 6(b); $t(14) = -1.91$, $p = 0.07$). Regarding the after-task phase, the parents continued commenting on the task in ADI, while they tended to suppress their body movement in IDI. Hence, in ADI their face became salient and attracted more attention, which shows that they did not focus on highlighting the final state of the cups but rather linguistically explained it.

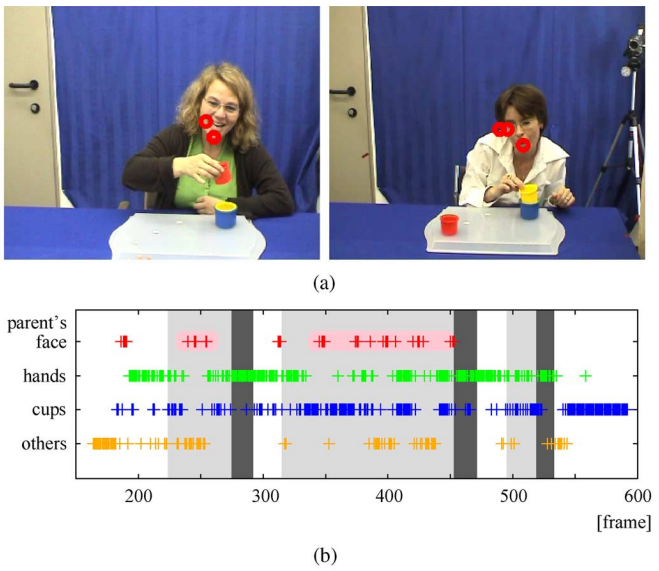

Fig. 8. Parental social feedback indicating significant event *beforehand* in IDI. (a) Talking to and smiling at infant before putting cup down into another. (b) Attention to mother's face before each significant event, corresponding to right mother in (a).

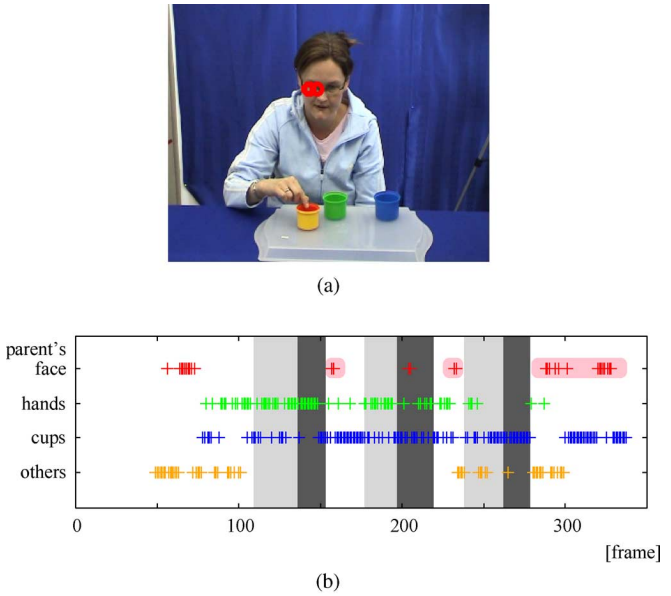

Fig. 9. Parental social feedback indicating significant event *afterward* in IDI. (a) Addressing infant after placing cup into another. (b) Attention to mother's face after each significant event.

C. Underlining Properties of Cups

We next analyzed how much the inherent features of the cups were highlighted by parental action. The inherent features here include the color, the intensity, and the orientation. It is considered that in action learning, not only the movement of the cups but also the properties of them should be attended to. For instance, what color of a cup is moved first and what size it is are important to appropriately stack the cups. Therefore, we focused on the properties of the cups and examined how the inherent features of the cups were underlined in terms of their contribution to the saliency.

Fig. 10 shows the contribution rate of the inherent features to the saliency for the cups. The rate was calculated only when the cups were attended to by the saliency model. As in Fig. 6, the results were compared between IDI and ADI in each task phase. The paired t-test revealed a significant IDI-ADI difference in the before-task phase (the leftmost in Fig. 10; $t(14) = 2.42$, $p < 0.05$) and also in the during-task phase (the second left in Fig. 10; $t(14) = 3.58$, $p < 0.05$). These results indicate that in the two phases, the properties of the cups were highlighted by parental action modification.

The reasons are considered as follows: As described in the former sections, in IDI the parents tended to pause their body movement before starting the task and even while demonstrating it. In the before-task phase, they took a long pause to emphasize the initial state of the cups. This behavior made the properties of the cups more visible. Similarly, the parents sometimes stopped their cup-handling movement during the task demonstration. To indicate the significant events beforehand and/or afterward, they paused their hands' movement and verbally addressed the infants. At that moment, they also closely presented the cups to the infants by suppressing their body movement. They even stopped talking to the infants to draw the infants' attention to the cups.

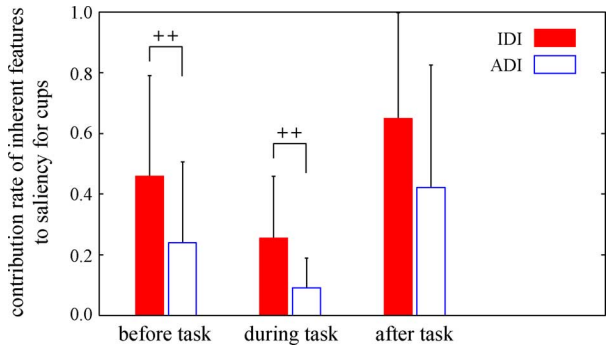

Fig. 10. Contribution rate of inherent features to saliency for cups (++: significant difference).

This behavior consequently made the properties of the cups contribute more to the saliency.

In ADI, by contrast, the properties of the cups were not visible over the task demonstration, but rather the motion features were. Although during the task, the cups attracted as much attention in ADI as in IDI [see the third left in Fig. 6(b)], their inherent features were not detectable to the saliency model. The reason is that in ADI, the parents kept moving to achieve the task and did not emphasize the cups' state. They might suppose that the adult partners could easily perceive the cups and thus did not need to closely observe them. This result shows how differently the parents assumed the perceptual ability of the partners.

VI. Discussion Toward Scaffolding Robot Action Learning

Our analysis employing the saliency model revealed that motionese can do the following:

- highlight the initial and final states of the objects involved in the task;

- provide frequent social signals indicating the significant events; and
- emphasize the properties of the objects.

For robots that learn to imitate and understand actions, it is important to extract the relevant features of the actions and presume the goal of them. This section first discusses how the effects brought by motionese can contribute to inferring the goal of the task and determining what to imitate. Then, we present our idea and ongoing work toward developing such an architecture.

A. Motionese Tells What to Imitate

First, highlighted initial and final states of objects enable robots to detect the state change in the objects and thus presume the goal of the task. In our experiment, the goal was to move the red, the yellow, and the green cups into the blue one. In IDI, the parents emphasized the cups before starting to demonstrate the task and after fulfilling it by suppressing their body movement and/or adding movement to the cups. Of particular note is that even in the latter case, i.e., when the parents tapped the first cup on the table, the position of the cup was still kept so that the initial state was visible. It therefore suggests that even if robots have no task knowledge, motionese can strongly support them in inferring the goal of the task.

Secondly, parental frequent social feedback can inform robots about the sub-goals of the action. The sub-goals of the stacking-cups task were to move one of the three cups into the blue one. Our analysis revealed that in IDI, the parents gave social signals shortly before and/or after achieving the each sub-goal. The preceding social signals allow robots to pay careful attention to the following event, while the signals after the event provide immediate feedback on the achieved action. Such immediate feedback has advantage over delayed feedback, which is often discussed as a problem in robot action learning [31], [32]. Although recognizing the meaning of social feedback is another challenge, motionese is suggested to help robots detect meaningful segments of the action.

Thirdly, we state that the underlined properties of the objects also contribute to presuming the goal of the task. As described in Section V-C, the inherent features of the objects impart more knowledge about the task, e.g., what size and color of a cup should be moved first in order to successfully stack the cups. From the qualitative analysis on the parental actions, we found that in IDI, the parents taught the properties of the cups separately from the movement. At each moment they seemed to focus either on presenting the cups or on showing the movement involving the cups. In ADI, such behavior was not found. The parents just kept demonstrating the movement and did not take time to introduce the cups. This difference suggests that motionese can also help robots schedule the learning phases for objects and movement.

Although these findings strongly support that motionese assists robots in learning the goal of the task, overcoming the issue of what to imitate requires a further step from detecting the goal: Robots have to know whether the goal is more important than the means or vice versa. The stacking-cups task is a goal-oriented, and thus the parents emphasized the state of the cups. In a means-oriented task, by contrast, parents might underline the movement of objects rather than the state. They would selectively modify their actions so that robots and infants can learn the trajectory of the movement as well as the goal of it. Our first analysis comparing a sprinkling-salt task (i.e., to get salt from a salt dispenser by tilting and/or tapping it, whose means is more crucial) to the stacking-cups task revealed some differences as well as similarities in parental actions depending on the task [33]. In order to strengthen our suggestion, we intend to more closely analyze our data and extend our analysis to diverse types of task demonstrations.

B. Designing Robots That Learn Action From Parental Demonstration

Designing a robot architecture to learn actions from parental demonstrations raises the following questions:

- How can robots induce parent-like teaching of human partners?
- How can robots understand social signals from human partners?
- How can robots associate the information extracted from human teaching?

An open question in human–robot interaction is how people want to teach robots, i.e., whether they accept robots as infant-like agents or adult-like. It is obvious that the behavior of robots as well as their appearance affects people's acceptance. Levin and colleagues have addressed the question focusing on the appearance [34], [35]. They found that people modified their teaching strategy for a computer learner presented in a picture as for an infant also in a picture. Our focus, in contrast, is on the robots' behavior. We suggest that robots' attention employing the saliency model can yield infant-likeness of the robots and thus induce parent-like teaching of human partners. Our experiment on human–robot interaction supported our hypothesis: Our robot simulation equipped with the saliency model elicited action modifications of partners as observed in parent-infant interaction [36], [37]. We will next investigate the synergistic effect of robots' behavior and appearance.

Incorporating a mechanism to comprehend social cues from human partners is also an interesting issue. From our analysis, we found out that social signals given by demonstrators play an important role in detecting subgoals of the task. Developmental studies, on the one hand, suggest that infants at 12 months are able to refer to the emotional expression of their parents and also to follow the direction of the parents' gaze [38]. In developmental robotics, on the other hand, learning models for joint attention have been actively investigated [39]–[42]. The researchers demonstrated that robots can learn to follow human gaze by detecting contingency between his/her face image and a salient object. We consider that integrating such a mechanism with the saliency model will enable robots to better understand what to imitate.

After extracting the important features of a demonstrated task, robots have to associate them in order to determine what to imitate and moreover how to imitate. For example, the saliency model has robots attend to the demonstrator's face when he/she is providing a social cue. It is, on the one hand, important to detect a significant event in the task, but on the other hand, robots have to disregard it when reproducing the task. Robots

do not need to imitate the demonstrator's facial expression. A possible idea to cope with the problem is to exploit continuity. Examining the temporal and spatial continuity of the extracted information enables robots to determine what are relevant to achieving the task and how they can be associated [43]. We will extend our architecture so as to address the issue of how to imitate beyond what to imitate.

VII. Conclusion

We investigated how motionese can assist robots in learning actions. A difficulty in robot action learning is that robots do not know to which aspects of the demonstrated action they should attend although exposed to a huge amount of sensory information. Inspired by parent-infant interaction, we hypothesized that motionese enables robots to detect the relevant features of the action. Parental action modification such as suppression and addition of their body movement physically emphasizes the important aspects of the action, so that they can draw bottom-up attention. Our analysis employing the saliency model revealed that motionese has the effect of highlighting the initial and final states of the action, the significant events in it, and the properties of the objects, which can impart the goal of the action.

Our future issue is to develop robots that can take advantage of human scaffolding. It involves designing human-robot interaction like parent-infant interaction as well as building robot learning models. Addressing the questions raised in the discussion will lead us to the goal.

References

[1] J. Call and M. Carpenter, "Three sources of information in social learning," in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. L. Nehaniv, Eds. : MIT Press, 2002, pp. 211–228.

[2] C. L. Nehaniv and K. Dautenhahn, "Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications," in *Interdisciplinary Approaches to Robot Learning, World Scientific Series in Robotics and Intelligent Systems*, J. Demiris and A. Birk, Eds. : , 2000, vol. 24.

[3] C. L. Nehaniv and K. Dautenhahn, "Like me? Measures of correspondence and imitation," *Cybern. Syst.: Int. J.*, vol. 32, pp. 11–51, 2001.

[4] C. Breazeal and B. Scassellati, "Challenges in building robots that imitate people," in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. L. Nehaniv, Eds. Cambridge, MA: MIT Press, 2002, pp. 363–389.

[5] C. Breazeal and B. Scassellati, "Robots that imitate humans," *Trends in Cognitive Sciences*, vol. 6, no. 11, pp. 481–487, 2002.

[6] M. Carpenter and J. Call, "The question of 'what to imitate': Inferring goals and intentions from demonstrations," in *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*, C. L. Nehaniv and K. Dautenhahn, Eds. Cambridge, U.K.: Cambridge University Press, 2007, pp. 135–151.

[7] 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.

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

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE 48th Annu. Int. Symp. Opt. Sci. Technol.*, 2003, pp. 64–78.

[11] A. Billard, Y. Epars, S. Calinon, S. Schaal, and G. Cheng, "Discovering optimal imitation strategies," *Robot. Autonom. Syst.*, vol. 47, pp. 69–77, 2004.

[12] A. G. Billard, S. Calinon, and F. Guenter, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robot. Autonom. Syst.*, vol. 54, no. 5, pp. 370–384, 2006.

[13] S. Calinon and A. Billard, "What is the teacher's role in robot programming by demonstration? – Toward benchmarks for improved learning," *Interact. Studies*, vol. 8, no. 3, pp. 441–464, 2007.

[14] S. Calinon and A. Billard, "A framework integrating statistical and social cues to teach a humanoid robot new skills," in *Proc. ICRA Workshop Social Interact. Intell. Indoor Robots*, 2008.

[15] A. Alissandrakis, C. L. Nehaniv, and K. Dautenhahn, "Imitation with alice: Learning to imitate corresponding actions across dissimilar embodiments," *IEEE Trans. Syst., Man, Cybern.-Part A: Syst. Humans*, vol. 32, no. 4, pp. 482–496, 2002.

[16] B. Scassellati, "Knowing what to imitate and knowing when you succeed," in *Proc. AISB'99 Symp. Imitat. Animals Artifacts*, 1999, pp. 105–113.

[17] R. J. Brand, W. L. Shallcross, M. G. Sabatos, and K. P. Massie, "Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infant-versus adult-directed action," *Infancy*, vol. 11, no. 2, pp. 203–214, 2007.

[18] R. J. Brand and W. L. Shallcross, "Infants prefer motionese to adult-directed action," *Develop. Sci.*, vol. 11, no. 6, pp. 853–861, 2008.

[19] N. Masataka, "Motherese in a signed language," *Infant Behav. Develop.*, vol. 15, pp. 453–460, 1992.

[20] N. Masataka, "Perception of motherese in a signed language by 6-month-old deaf infants," *Develop. Psychol.*, vol. 32, no. 5, pp. 874–879, 1996.

[21] N. Masataka, "Perception of motherese in japanese sign language by 6-month-old hearing infants," *Develop. Psychol.*, vol. 34, no. 2, pp. 241–246, 1998.

[22] P. Zukow-Goldring, "Assisted imitation: Affordances, effectivities, and the mirror system in early language development," in *Action to Language Via the Mirror Neuron System*, M. A. Arbib, Ed. Cambridge, U.K.: Cambridge University Press, 2006, pp. 469–500.

[23] P. Zukow-Goldring and M. A. Arbib, "Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention," *Neurocomputing*, vol. 70, no. 13–15, pp. 2181–2193, 2007.

[24] C. Yu, L. B. Smith, and A. F. Pereira, "Embodied solution: The world from a toddler's view," in *Proc. 7th IEEE Int. Conf. Develop. Learning*, 2008.

[25] J. Fritsch, N. Hofemann, and K. Rohlfing, "Detecting 'when to imitate' in a social context with a human caregiver," in *Proc. ICRA Workshop Social Mechan. Robot Programming by Demonstr.*, 2005.

[26] J. Schmidt, J. Fritsch, and B. Kwolek, "Kernel particle filter for real-time 3d body tracking in monocular color images," in *Proc. Autom. Face Gesture Recogn.*, 2006, pp. 567–572.

[27] R. M. Golinkoff and K. Hirsh-Pasek, "Baby wordsmith: From associationist to social sophisticate," *Current Directions Psychol. Sci.*, vol. 15, no. 1, pp. 30–33, 2006.

[28] M. Schlesinger, D. Amso, and S. P. Johnson, "Simulating infants' gaze patterns during the development of perceptual completion," in *Proc. 7th Int. Conf. Epigenetic Robot.*, 2007, pp. 157–164.

[29] 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.

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

[31] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[32] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda, "Purposive behavior acquisition for a real robot by vision-based reinforcement learning," *Machine Learning*, vol. 23, pp. 279–303, 1996.

[33] Y. Nagai and K. J. Rohlfing, "Parental action modification highlighting the goal versus the means," in *Proc. IEEE 7th Int. Conf. Develop. Learning*, 2008.

[34] S. Killingsworth, M. M. Saylor, and D. T. Levin, "Segmenting action for computers and humans: Possible links to intentional understanding," in *Proc. 2005 IEEE Int. Workshop Robot Human Interactive Commun.*, 2005, pp. 196–201.

[35] J. S. Herberg, M. M. Saylor, D. T. Levin, P. Ratanaswasd, and D. M. Wilkes, "The perceived intentionality of an audience influences action demonstrations," in *Proc. 5th Int. Conf. Develop. Learning*, 2006.

[36] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *Proc. 2008 IEEE Int. Conf. Robot. Automat.*, 2008, pp. 3545–3550.

[37] C. Muhl and Y. Nagai, "Does disturbance discourage people from communicating with a robot?," in *Proc. 16th IEEE Int. Symp. Robot Human Interactive Commun.*, 2007, pp. 1137–1142.

[38] C. Moore and P. J. Dunham, *Joint Attention: Its Origins and Role in Development*. : Lawrence Erlbaum Assoc Inc, 1995.
[39] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Sci.*, vol. 15, no. 4, pp. 211–229, 2003.
[40] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Adv. Robot.*, vol. 20, no. 10, pp. 1165–1181, 2006.
[41] J. Triesch, C. Teuscher, G. O. Deak, and E. Carlson, "Gaze following: Why (not) learn it?," *Develop. Sci.*, vol. 9, no. 2, pp. 125–147, 2006.
[42] B. Scassellati, "Theory of mind for a humanoid robot," *Autonom. Robots*, vol. 12, pp. 13–24, 2002.
[43] Y. Nagai, "From bottom-up visual attention to robot action learning," in *Proc. 8th IEEE Int. Conf. Develop. Learning*, 2009, to be published.

Yukie Nagai received the Master degree in engineering from Aoyama Gakuin University, Japan in 1999, and the Ph.D. degree in engineering from Osaka University, Japan, in 2004.

From 2002 to 2004, she was a Research Associate at the Graduate School of Engineering, Osaka University. From 2004 to 2006, she was a Researcher at the National Institute of Information and Communications Technology, Japan. Since 2006, she has been a member of the Applied Computer Science, Faculty of Technology, Bielefeld University, Germany. In 2008, she also joined the Research Institute for Cognition and Robotics, Bielefeld University. She has been working in the research field of cognitive developmental robotics. Her research interests are the developmental mechanisms for social abilities, e.g., joint attention, imitation, and language use, in robots as well as in human infants. Her aim is at understanding human cognition by modeling and evaluating artificial systems. She has been investigating how caregivers can scaffold infants' and robots' learning while they induce the scaffolding.

Katharina J. Rohlfing received the Master's degree in linguistics, philosophy, and media Studies from the University of Paderborn, Germany, in 1997. From 1999 to 2002, she was a member of the Graduate Program Task Oriented Communication. She received the Ph.D. degree in linguistics from the Bielefeld University, Germany, in 2002. Her postdoctoral work at the San Diego State University, the University of Chicago, and Northwestern University was supported by a fellowship within the Postdoc-program of the German Academic Exchange Service (DAAD) and by the Emmy Noether-Programme of the German Research Foundation (DFG).

In 2006, she became a Dilthey-Fellow (Funding initative "Focus on the Humanities"), and her research is supported by the Volkswagen Foundation. Since May 2008, she has been Head of the Emergentist Semantics Group within Center of Excellence Cognitive Interaction Technology, Bielefeld University. She is interested in learning processes and in her research, she has been investigating the interface between early stages of language acquisition and conceptual development.