Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations?

Katrin Solveig Lohan Anna-Lisa Vollmer Jannik Fritsch Katharina Rohlfing
Britta Wrede
CoR-Lab, Applied Informatics Group
Bielefeld University, Bielefeld, Germany
http://www.cor-lab.de
klohan@techfak.uni-bielefeld.de

Abstract

In developmental research, tutoring behavior has been identified as scaffolding infants' learning processes. Infants seem sensitive to tutoring situations and they detect these by ostensive cues [4]. Some social signals such as eye-gaze, child-directed speech (Motherese), child-directed motion (Motionese), and contingency have been shown to serve as ostensive cues. The concept of contingency describes exchanges in which two agents interact with each other reciprocally. Csibra and Gergely argued that contingency is a characteristic ostensive stimulus of a tutoring situation [4]. In order for a robot to be treated similar to an infant, it has to both, be sensitive to the ostensive stimuli on the one hand and induce tutoring behavior by its feedback about its capabilities on the other hand.
In this paper, we raise the question whether a robot can be treated similar to an infant in an interaction. We present results concerning the acceptance of a robotic agent in a social learning scenario, which we obtained via comparison to interactions with 8-11 months old infants and adults in equal conditions. We applied measurements for motion modifications (Motionese) and eye-gaze behavior. Our results reveal significant differences between Adult-Child Interaction (ACI), Adult-Adult Interaction (AAI) and Adult-Robot Interaction (ARI) suggesting that in ARI, robot-directed tutoring behavior is even more accentuated in terms of Motionese, but contingent responsivity is impaired. Our results confirm previous findings [14] concerning the differences between ACI, AAI, and ARI and constitute an important empirical basis for making use of ostensive stimuli as social signals for tutoring behavior in social robotics.

1. Introduction

In social learning, infants benefit from the behavior of their tutors. The modified behavior seems to help infants to filter the information that is crucial for learning. Csibra and Gergely [4] highlight the importance of this pedagogic behavior that is crucial for the understanding of some actions: "pedagogy essentially created a new way of information transfer among individuals through the use of ostensive communication". In their work, they give the example of peeling a hard fruit or carve away pieces of wood with a tool. The movement and the tool in both actions are the same, but the goal and reason for the action are very different. Where it is easy to infer the goal of the action when peeling a fruit, i.e. getting to the edible parts, it is not obvious what is intended in the case of the wood carving. Therefore, tutoring is crucial in order for a learner to understand the goal correctly. Csibra and Gergely [4] argue that economical reasons account for tutoring, because otherwise learning would not be feasible. Tutoring situations thus are created by the tutor via ostensive stimuli, which are "originally evolved to assist pedagogy". The effect of pedagogy seems to rely on the bidirectionality. Csibra and Gergely (2005) explain the contribution achieved by the learner, who has to send signals during the course of tutoring telling the tutor when s/he is attentive and receptive and possibly showing understanding. Furthermore, infants seem sensitive to tutoring situations and ostensive cues help them to detect these [13]. The term "ostensive cues" refers to social signals such as eye-gaze, child-directed speech (Motherese) [5], child-directed motion (Motionese) [2,6,7], and contingency [4]. While the phenomenon of multimodal child-directed speech (Motherese) or action (Motionese) is widely known, the concept of contingency is less popular. It describes exchanges in which two agents interact with each other reciprocally. Csibra and Gergely ([4], p.8) argue that contingent responsivity is a characteristic ostensive stimulus of a tutoring situation: "If a source repeatedly appears to remain silent during your actions but starts to emit signals as soon as you have stopped your actions, it gives

978-1-4244-4799-2/09/$25.00

you the strong impression that the source is communicating with you". The idea of creating a robot that actively filters the information from the environment and manages to attend to certain sources of information while ignoring others has to be supported by the robot's sensitivity to the ostensive stimuli on the one hand and induce tutoring behavior by its feedback about its capabilities on the other hand. A robot which has the appearance of an infant should hence be able to profit from these behavior modifications as well. Recently, Vollmer et al. found that adults modify their behavior when interacting with children (ACI) and robots (ARI) as opposed to adult-directed interaction (AAI) [14]. Modifications were found with respect to Motionese measurements, indicating that in ACI and ARI movements were slower, less round and had a slower pace than in AAI indicating that subjects behave similar towards robots and infants. However, number and length of eye-gaze bouts differed significantly between ACI and ARI with less eye-gaze bouts and less long eye-gaze bouts directed towards the interaction partner in ARI. This indicates that contingency was impaired in the ARI condition. In this paper, we report on results from a task with a similar structure based on a more fine grained analysis of the eye-gaze behavior in order to

- show how far the findings by Vollmer et al. hold for a different task
- analyze the structure of eye-gaze behavior over time and
- discuss these results with respect to the question in how far the observed modifications of behavior can be interpreted as ostensive signals in human-robot interaction.

2. Experiment

Two experiments were carried out to obtain data from parent-infant and adult-robot interactions [14]. The data on adult-child interaction is based on the same setting as in [12] and [10]. The data on human-robot interaction was obtained in a second experiment as described in [14]. From the overall set of items that were presented we selected the "Minihausen" task. This task is similar to the stacking-cups task as it is a rather goal-directed action with three sub-goals to be reached. Results from analyses of motionese and contingency features in parent-infant and adult-robot interaction have shown that while motionese features of infant-directed and robot-directed interactions are similar, they diverge for contingency measures, indicating that contingency is impaired in human- robot-interaction, [14]. In this paper we ask the question in how far these results are decisive for the statement that motionese as well as contingency features serve the function of ostensive signals.

2.1. Motionese Experiment (ME)

2.1.1 Subjects

The Motionese Corpus consists of infant- and adult-directed interactions. We selected the younger group comprising 12 families of 8 to 11 months old children. Both parents were asked to demonstrate functions of 10 different objects to their children as well as to their partners or another adult. In the following, we focus on the analysis of the "Minihausen" task, because it offers good comparability in motion performance. We further selected a subgroup of 8 parents (4 fathers and 4 mothers) for the ACI and a subgroup of 12 parents (7 fathers and 5 mothers) for the AAI, because of the quality of the video, sound and due to the way in which the action was performed. More specifically, the order in which the blocks of the considered "Minihausen" task are put onto the wooden base poles can vary: We selected only those parents, who started the task by putting the first block -the one closest to the body- onto the respective pole which means putting the blue block onto the rightmost pole. (see Fig. 3 a1).

2.1.2 Setting

Parents were instructed to demonstrate a "Minihausen" task to an interaction partner. The interaction partner was first their infant and then an adult. Fig. 1 illustrates the top-view of the experimental setup. The "Minihausen" task was to sequentially pick up the blue (a1), the yellow (a2), and the green (a3) block and put them onto the wooden base with three poles on the white tray.

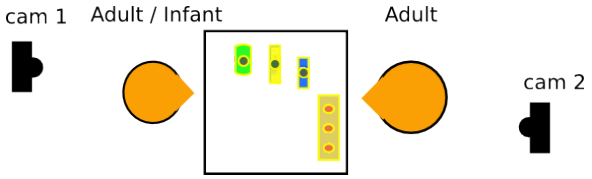

Figure 1. Motionese Setting, there are two cameras which are recording the scene. The interaction partners are seated across from each other and the object is laid on the table in front of the tutor.

2.2. Robot-Directed Interaction Experiment (RDIE)

2.2.1 Subjects

31 adults (14 females and 17 male) participated in this experiment 7 out of which were parents as well. Out of this group, we selected 12 participants (8 female and 4 male), who performed the task in a comparable manner.

2.2.2 Setting

The participants were instructed to demonstrate several objects to an interaction partner, while explaining him/her how to do it (Fig. 2). Again we chose the "Minihausen" task for analysis. The interaction partner was an infant-like looking virtual robot with a saliency-based visual attention system [10]. The robot-eyes will follow the most salient point in the scene, which is computed by color, movement, and other features (see [10]).

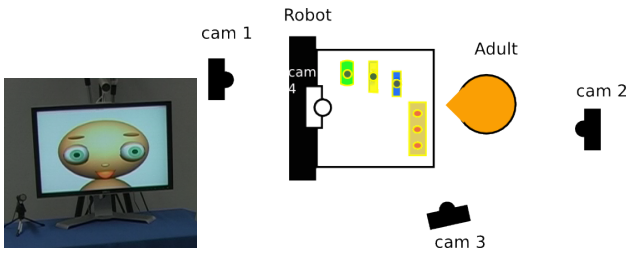

Figure 2. The robot simulation presented on the screen can be seen on the left picture. The rigth picture shows the Robot-directed Interaction Setting, there are four cameras which are recording the scene. The subject is seated across from the robot and the object is laid on the table in front of the tutor.

3. Data Analysis

The goal of this paper was to analyze those cues, that we hypothesize to serve as social signals in tutoring behavior. These can be grouped into two groups, one that measures Motionese and another one that that may be used to measure Contingency. We coded the videos semi-automatically to obtain data for the 2D hand trajectories and the eye gaze directions.

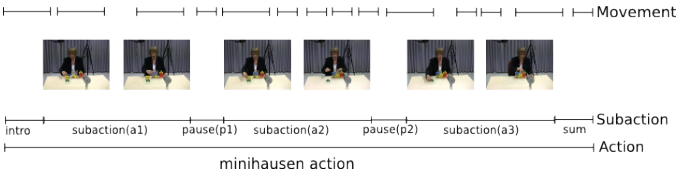

Figure 3. The action was devided into movment and pause parts and into subactions. This graphic shows an example for the structure of an 'Action', 'Subaction'(intro = Introduction and sum = summary), and 'Movement'.

3.1. Annotations

For all annotations, we used the video captured by camera (cam) 1, see Fig. 1 and 2. It shows the front view on the demonstrator and is therefore best suited for action, movement, and gaze annotations, which are discussed in detail below.

3.1.1 Motionese

Action Segmentation: For analyzing the data, the action of the "Minihausen" task and additionally, the sub-actions (a1-a3) of grasping one block until releasing it onto the end position (Fig. 3) were marked in the video. We defined

1. action as the whole process of transporting all objects to their goal positions.

2. subaction as the process of transporting one object to its goal position.

3. movement as phases where the velocity of the hand is above a certain threshold. All other phases are defined as pauses.

Hand Trajectories: The videos of the two experiments were analyzed via a semiautomatic hand tracker system (Fig. 4). The system is written as a plug-in for a graphical plug in shell, iceWing [8], and makes it possible to track both hands with an Optical Flow based algorithm, Lucas & Kanade [9]. It allows manual adjustment in case of tracking deviation. We used this tracking system instead of a previously used 3D body model system, [12], since 3D results in [12] were not significant, we focused on 2D analyses which provide to show more stable results. Additionally, the new system is easily accessible for non-expert users.

Figure 4. Example frame for hand tracker system annotation. The red and violet circles depict the tracking regions. The points in the middle of the circles are the resulting 2D points for the hand trajectory.

3.1.2 Contingency

Eye Gaze: In annotating the eye gaze directions with the program Interact [1], we distinguished between looking at the interaction partner, looking at the object and looking anywhere else.

3.2. Measures

For quantifying Motionese and Contingency, we computed five variables related to the 2D hand trajectories derived from the videos and the eye gaze bout annotations produced with Interact.

3.2.1 Motionese

We measured Motionese in terms of velocity and range as defined in [14].

Velocity was computed using the derivative of the 2-dimensional hand coordinates of the hand which performed

the action per frame as the average velocity for subactions a1, a2, and a3 each.

Range was defined for each subaction separately as the covered motion path divided by the distance between motion, i.e. subaction, on- and offset.

3.2.2 Contingency

The Contingency of the interactions was quantified in terms of variables related to eye gaze, as defined in [3] for measuring interactiveness.

The *total length of eye-gaze bouts to interaction partner* defined as the percentage of time of the action spent gazing at the interaction partner was computed. Brand et al. found that the total length of eye-gaze bouts to the interaction partner in their study was significantly greater in ACI than in AAI [3]. Also the *total length of eye-gaze bouts to object* and the *total length of eye-gaze bouts elsewhere* were calculated as the percentage of time of the action spent gazing at the object and somewhere else, as for example at the table or the experimentor.

4. Results

A non-parametric test (Mann-Whitney U test) was run for all pairs of samples, ACI vs. AAI, ACI vs. ARI, and AAI vs. ARI. Table 1 depicts the results of the study.

4.1. Motionese

For the Motionese measures, our results revealed the following:

For the *velocity* measure, which is computed for each subaction and takes into account the hand movement during the transportation of the respective block, the results showed significant differences for all three subactions for all pairs of conditions. These results clearly show that in AAI hand movements are faster than in ACI and ARI and additionally that hand movement is slowest in the ARI condition. Also note that for all conditions the mean values increase for the consecutive subactions: velocity in subaction a1 $<$ velocity in a2 $<$ velocity in a3. In ARI, the rate in which the mean values increase is lowest and in AAI the rate is highest. The latter is specially noticable for the last subaction a3.

The *range* measure suggests that ARI exhibits the greatest range for each subaction and therefore movement is most exaggerated. Also, range is greater in ACI than in AAI. For ACI vs. AAI results revealed no significance, but a trend for subactions a2 and a3. For ACI vs. ARI solely results for subaction a3 showed significance, for a1 and a2 they show a trend. For AAI vs. ARI subactions a2 and a3 revealed significance, whereas a1 again shows a trend. Again we can state that in ARI the first subaction a1 has the highest range value of all subactions over all conditions. Looking at this measure over time, range decreases rapidly to about one half for subaction a2 and some more for the last subaction a3. For the other conditions however the rate of change, i.e. the decrease, is not as drastical.

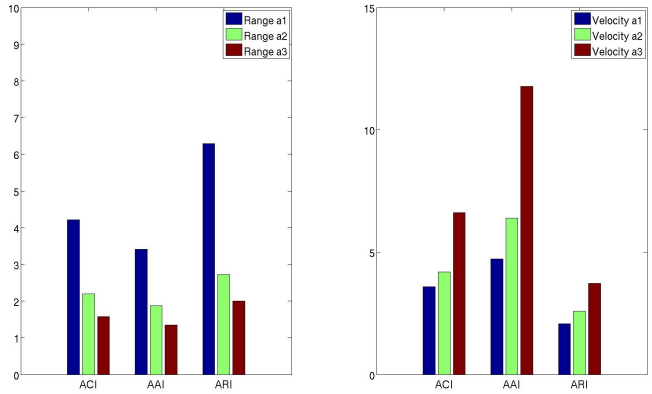

Figure 5. This graph shows the range of hand movement in the three different subactions on the left. On the right, the mean velocity of hand movement in the three different subactions can be seen for the "Minihausen"-task (y-axis) in every condition (x-axis).

4.2. Contingency

Most interestingly, the results for eye gaze show a completely different picture. For *total length of eye-gaze bouts to interaction partner* they show that in ACI significantly more time was spent gazing at the interaction partner than in AAI and ARI. Differences between AAI and ARI are not significant. Looking at this measure over time, it is interesting to notice that in all three conditions the most time of gazing at the interaction partner was spent in the summary part of the action, sum.

For the measure *total length of eye-gaze bouts to object*, values are significantly lower in ACI than in AAI and ARI, where differences between AAI and ARI exhibit that values are significantly lower in ARI.

The *total length of eye-gaze bouts elsewhere*, which measures the percentage of time gazed neither to interaction partner nor object, reveals that most time gazing somewhere else is spent in the ARI condition, followed by ACI. The differences between ACI and AAI could be a result of the design of the study, because the AAI follows the ACI, so that instructions and experimentor are not anymore needed to turn to for help in the demonstration of the task, because it has already been shown once. Additionally, in all conditions it is gazed elsewhere mostly in p1 and p2 and not during the transportation of the cups in a1, a2 and a3.

5. Conclusion

To conclude, we did find ostensive signals in tutoring situations in adult-robot interaction. On the one hand, our results for range and velocity show significantly exaggerated

Variable	ACI		AAI		ARI		ACI vs AAI	ACI vs ARI	AAI vs ARI
	M	*SD*	*M*	*SD*	*M*	*SD*	*Z*	*Z*	*Z*
velocity a1	3.58	0.81	4.72	1.39	2.08	0.86	−2.394**	−3.668***	−3.747***
velocity a2	4.19	1.84	6.39	1.71	2.59	0.87	−2.535**	−2.792**	−3.982***
velocity a3	6.62	2.43	11.78	2.95	3.73	1.51	−3.098***	−2.956**	−3.982***
range a1	4.22	2.49	3.41	0.72	6.29	5.53	−0.211	−1.369+	−1.288+
range a2	2.19	0.48	1.88	0.25	2.72	0.97	−1.549+	−1.314+	−2.635**
range a3	1.57	0.37	1.35	0.09	2	0.56	−1.479+	−2.409**	−3.396***
total length eye-gaze to i.p. in	10.86	14.52	6.65	7.15	6.65	7.15	−0.833	−1.419+	−0.76
total length eye-gaze to i.p. a1	27.81	25.02	9.01	16.92	9.25	11.38	−2.2*	−1.882*	−0.97
total length eye-gaze to i.p. p1	24.19	28.17	3.7	9.71	7.35	8.78	−1.853*	−1.03	−1.634+
total length eye-gaze to i.p. a2	15.39	16.67	2.42	4.44	3.16	4.81	−2.054*	−2.066*	−0.244
total length eye-gaze to i.p. p2	33.73	24.63	2.61	7.09	2.69	5.9	−3.055***	−3.306***	−0.082
total length eye-gaze to i.p. a3	23.05	23.09	4.37	8.71	6.2	10.48	−2.273*	−2.292*	−0.384
total length eye-gaze to i.p. su	43.8	23.81	27.55	7.43	19.66	13.65	−0.493	−2.793**	−1.878+
total length eye-gaze to o. in	69.29	29.43	82.32	22.47	62.65	8.7	−1.353+	−1.15	−2.817**
total length eye-gaze to o. a1	70.94	22.72	89.52	16.69	83.21	13.46	−2.1*	−1.213	−1.155
total length eye-gaze to o. p1	60.95	26.97	88.99	23.87	68.36	25.95	−2.273*	−0.714	−2.097*
total length eye-gaze to o. a2	82.68	18.18	96.2	8.19	92.43	7.85	−2.198*	−1.308+	−1.533+
total length eye-gaze to o. p2	65.02	25.55	97.39	7.09	80.23	22.36	−3.055***	−1.503+	−2.092*
total length eye-gaze to o. a3	76.95	23.25	95.63	8.71	87.23	13.77	−2.273*	−1.252	−1.721*
total length eye-gaze to o. su	55.79	22.63	52.71	31.88	57.92	17.94	−0.352	−0.109	−0.527
total length eye-gaze e. in	20.89	29.12	11.03	18.15	34.93	9	−0.624	−1.984*	−3.127***
total length eye-gaze e. a1	1.91	4.75	1.48	4.67	7.53	10.61	−0.52	−1.625+	−1.919*
total length eye-gaze e. p1	16.09	19.93	7.32	23.14	24.29	26.94	−1.501+	−0.812	−1.952*
total length eye-gaze e. a2	2.51	3.9	1.37	4.34	4.41	7.42	−1.178	−0.371	−1.604+
total length eye-gaze e. p2	2.38	5.35	0	0	17.08	20.59	−1.382+	−1.879*	−2.551**
total length eye-gaze e. a3	0.74	1.67	0	0	6.57	12.94	−1.382+	−0.877	−1.803*
total length eye-gaze e. su	1.09	2.31	7.65	11.74	22.42	15.92	−1.091	−3.507***	−2.267*

Table 1. Results of Mean, Standard deviation, Mann-Whitney U test, $+p$ <0.1, $*p$ <0.05, $**p$ <0.01, $***p$ <0.001, interaction partner ($i.p.$), object ($o.$)., else ($e.$). su = sum = summary, in = intro = introduction

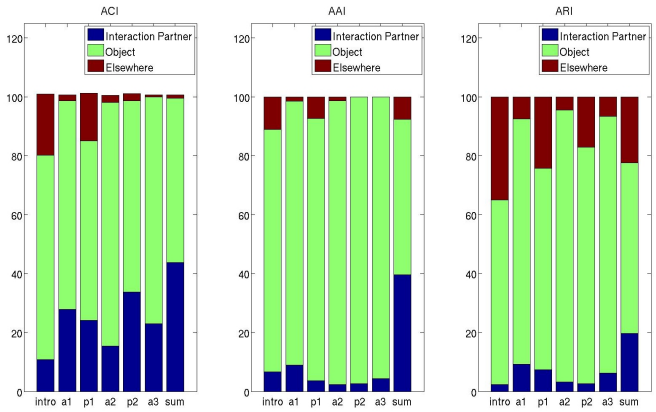

Figure 6. This graph shows the total length of eye-gaze bouts to the interaction partner, the object and somewhere else (y-axis) over time: all seven action parts are displayed (x-axis) for ACI (left), AAI (middle) and ARI (right) condition.

hand movements which are clearly distinguishable from those observable in adult-adult interactions and which are even more accentuated than the hand movements in child-directed tutoring. Thus, ostensive stimuli are present in robot tutoring. These however change over time as we have seen: range of motion decreases drastically, whereas velocity increases slowly. We therefore hypothesize that the reson for this lies in the behavior of the learner which shapes the behavior of the tutor as stated for eye gaze behavior and hand movements by Pitsch et al. [11]. This process could be interpreted as an alignment process where the tutor starts of by clearly signaling his intention of tutoring the infant. This signal decreases during the ongoing interaction while the tutor captures the infant's attention and while observing an understanding process in the infant. The final behavior may thus be described as consisting of fragmentary cues rather than the complete and exaggerated signal. On the other hand, our results reveal that in order to create a contingent interaction with the partner, the learner needs to produce a suitable feedback. This means that although the tutor's hand movements in robot-directed tutoring seem to be even slower and less round than in child-directed tutoring, the tutor's eyegazing behavior in robot-directed tutoring is suggestive of a lack of appropriate social signals on

the recipient's side: The percentage of time the interaction partner is viewed by the tutor is much lower in ARI than in ACI.

The ostensive signals considered here appear practical for the robot to detect situations in which it is being tutored, but we argue that a robot cannot make use of an important ostensive stimulus such as contingency without providing the "right" signals for the interactional construct. In detail, we find that already from the introduction on: the eye-gaze behavior in the ARI situation is rather similar to that of the AAI situation, with less time of the eye-gaze being spent on the interaction partner. This is congruent with previous findings from [14]. If we hypothesize that eye-gaze is also being used in order to check for understanding of the partner, the eye-gaze behavior directly after the end of a subaction becomes relevant. Indeed, we can see that the eye-gaze lengths in both pauses p1 and p2 are significantly longer in ACI as opposed to AAI. Thus, the parents appear to look for understanding in their infants. Interestingly, the behavior in ARI tends to be similar to the one in AAI indicating that adults behave differently towards robots. However, in p1 we see a trend for the eye-gaze lengths to be significantly longer in ARI as opposed to AAI. This might indicate that the subjects are looking out for signs of understanding in the robot as well. Yet, this behavior dramatically changes in p2 where the eye-gaze length is again decreased to the level of AAI, whereas it is even slightly increased in ACI. This may be interpreted as a reaction to missing signals of understanding from the robot. In the summary part of the action (sum), finally, the overall eye-gaze length towards the robot becomes significantly shorter than in both, ACI and AAI.

In order to confirm these results and our interpretation we are planning to carry out analyses of the joint eye-gaze behavior. We hypothesize that the robot is not able to establish mutual gaze especially in the pauses which then leads to the increase of eye-gaze towards the robot.

6. Outlook

These findings suggest that ostensive signals are present in human-robot tutoring situations and may be used for the robot to learn. However, in order for the robot to elicit a contingent interaction, it needs to provide ostensive signals that indicate its understanding. Based on our observations of the infants' behavior, these ostensive signals have to pertain to attention. That is, the robot has to provide eye gaze that signals attention and establishes joint attention as well as shared attention. Another behavior of the infants that was not modeled in the ARI condition was their attempts to reach and grasp the demonstrated objects. Further analyses need to be carried out in order to reveal the pattern of these reaching gestures - first impressions of the data suggest that they are far from random but only appear at the end of the demonstrated actions. If this is true, the reaching gestures could be interpreted as a signal that the infant has understood the goal of the action, or at least, the end of the action. Further signals which can be observed from the infants are facial expressions. Again, systematic analyses need to be carried out, but first impressions suggest that emotional feedback indicates affective reactions to the objects themselves, but also to the attention grabbing behavior of the tutor, and the reaching of the goal.

References

[1] http://www.mangold-international.com/en/products/interact.html.

[2] R. Brand, D. Baldwin, and L. Ashburn. Evidence for'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1):72–83, 2002.

[3] R. Brand, W. Shallcross, M. Sabatos, and K. Massie. Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infant-versus adult-directed action. *INFANCY*, 11(2):203–214, 2007.

[4] G. Csibra and G. Gergely. Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance*, 21, 2005.

[5] A. Fernald and C. Mazzie. Prosody and focus in speech to infants and adults. *Developmental Psychology*, 27(2):209–21, 1991.

[6] L. Gogate, L. Bahrick, and J. Watson. A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4):878–894, 2000.

[7] J. Iverson, O. Capirci, E. Longobardi, and M. Cristina Caselli. Gesturing in mother-child interactions. *Cognitive Development*, 14(1):57–75, 1999.

[8] F. Loemker, 2007. `http://icewing.sourceforge.net`.

[9] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International joint conference on artificial intelligence*, volume 81, pages 674–679, 1981.

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

[11] K. Pitsch, A. Vollmer, J. Fritsch, B. Wrede, K. Rohlfing, and G. Sagerer. On the loop of action modification and the recipients gaze in adult-child-interaction.

[12] K. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199, 2006.

[13] A. Senju and G. Csibra. Gaze following in human infants depends on communicative signals. *Current Biology*, 18(9):668–671, 2008.

[14] A. Vollmer, K. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. Rohlfing, and B. Wrede. People modify their tutoring behavior in robot-directed interaction for action learning. In *Proceedings of the International Conference on Development and Learning*, 2009.