Acquisition of Human–Robot Joint Attention through Real-time Natural Interaction

Koh Hosoda*,† Hidenobu Sumioka,* Akio Morita,* and Minoru Asada*†

*Department of Adaptive Machine Systems,
†HANDAI Frontier Research Center,
Graduate school of Engineering, Osaka University
hosoda@er.ams.eng.osaka-u.ac.jp, sumioka@er.ams.eng.osaka-u.ac.jp,
morita@er.ams.eng.osaka-u.ac.jp, asada@ams.eng.osaka-u.ac.jp

Abstract

Joint attention, a process to attend to the object that the other attends to is supposed to be important for human–robot communication as well as for human–human communication. We propose an architecture for acquiring joint attention within a certain time period for realizing natural human–robot interaction. The architecture has two featured modules: a self-organizing map that makes the leaning time shorter and an automatic visual attention selector that let the agent communicate with a human asynchronously. We implemented the proposed architecture in a real robot agent and found that 30 *minutes was enough for acquiring joint attention with two objects. We can conclude from preliminary experiments that even if the gaze preference of the robot is different from that of the human caregiver, it can acquire joint attention.*

1 Introduction

Recently, tasks required for a robot become difficult and complicated, and it is nearly impossible to program overall procedures by hand. Communicating with the robot through interaction will be one of solutions to avoid such a catastrophe. It may enable us to program the robot intuitively. Another fact to support the importance of communication is that non-robot-experts even use robot platforms since robot agents are now going out of the laboratories and factories. If the communication is properly designed, non-experts can use such robots in a natural way without having special knowledge on programming the robots.

In the human–human communication, the ability to attend to an object which someone else is attending to is important. Such process is called joint attention [1] and is supposed to be a basic element for other social cognitive functions such as language communication and mind reading [2, 3, 4]. In the human–robot communication, the robot's ability of joint attention is often explicitly pre-programmed by the designer [5, 6, 7, 8]. However, it is not argued how the robot can acquire such an ability of joint attention through interactions with its environment. Such an acquisition process through interaction between the agent and a human caregiver is recently studied intensively [9, 10, 11].

If we study on a learning agent interacting with a human, the learning time should be within a certain time period. Long learning time makes the attending strategy of the human, such as the preference change to saliencies and the frequency of gaze change, different from that in human–human interaction. In other words, the required learning time should be almost the same as that of a human since he/she also changes the behavior by observing the robot behavior. Therefore, we should study on the learning architecture that enables real-time learning. However, the importance of real-time communication between a human and a robot is not so far pointed out clearly in the context of joint attention to the best of the authors' knowledge.

In this paper, emphasizing on the importance of real-time interaction between a robot and a human caregiver, we propose an architecture for a robot to acquire the joint attention behavior within a reasonable period of time. One idea to realize such an architecture is to use a self-organizing map to compress the high-dimensional image into certain size, which makes the leaning time shorter. The other is to use automatic visual attention selector, which let the agent communicate with a human asynchronously. We implemented the proposed architecture in a real robot agent and found that 30 minutes was enough for acquiring joint attention with two objects. We can conclude from preliminary experiments that even if the gaze preference of the robot is different from that of the human caregiver, it can acquire joint attention.

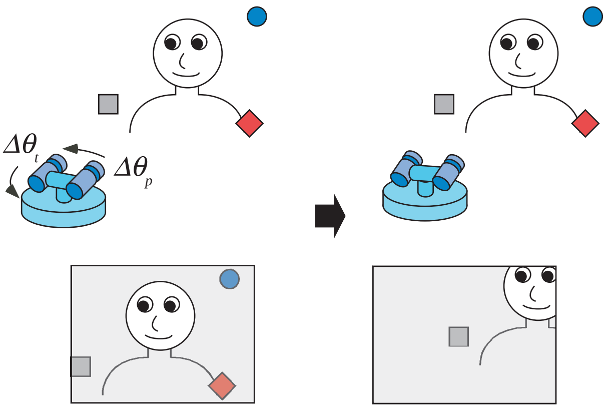

Figure 1: *Joint attention between a robot and a human caregiver: The robot observes the caregiver's face, and calculates head movement $\Delta\theta_t$ and $\Delta\theta_p$ so as to attend to the object that the caregiver attends to.*

2 Joint attention between a human caregiver and a robot

2.1 The joint attention behavior of the robot

The joint attention behavior of the robot with a human caregiver is shown in Figure 1. The robot observes the caregiver's face, and moves its head ($\Delta\theta_t$ and $\Delta\theta_p$ to the tilt and pan directions, respectively) to see the object that the caregiver is attending to. The robot acquires the relation between a face pattern of the caregiver and joint displacements to realize such a joint attention behavior through its experience. If there is only one object to be attended in the environment, the robot should only learn the relation between the face pattern and the movement generated by visual feedback to attend to the object. However, since there are several salient objects in the environment, the robot has to select one object to attend to.

2.2 Attending strategy of the human caregiver

The real environment has many objects that provide salient image features to the agents. Therefore, the attentional mechanism, how to select an object to attend to, should be a key to acquire the joint attention behavior. If the attention mechanisms of a human and of a robot are completely different, the behavior cannot be acquired. If the mechanisms are the same, the acquisition is easy. Then, the problem is to what extent the mechanisms share common elements.

Imagine the learning process of joint attention. If the communication is one way, that is, the caregiver looks what he/she wants to see while the robot learns the joint attention, the behavior of the robot does not change caregiver's attending strategy such as the preference change to saliencies and the frequency of gaze change. However, real communication is bilateral, that is, the caregiver sees the behavior of the robot and changes the strategy. The caregiver may be able to gradually change the attention strategy by observing the robot' behavior, and as a result, even if the shared common parts are small at the beginning of learning, eventually the joint attention behavior can be learned. In this sense, we cannot separate the learning of the agent from caregiver's strategy change.

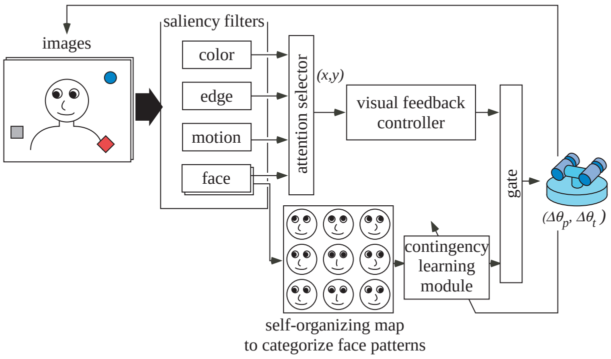
images
saliency filters
color
edge
motion
face
attention selector
(x,y)
visual feedback controller
gate
contingency learning module
self-organizing map to categorize face patterns
$(\Delta\theta_p, \Delta\theta_t)$

Figure 2: *An architecture for emergence of joint attention through natural interaction: A saliency filer extracts saliencies from the captured images. A attention selector will select one of saliencies and feeds its coordinates to the visual feedback controller. At the beginning of learning, the gate uses visual feedback. Meanwhile, the face pattern of the human caregiver is categorized by the self-organizing map. When the robot success to see the salient object in the center of the view, it will strengthen the connection between the face pattern and joint displacements. Over time, the gate gradually use the output of the learning module more than the visual feedback controller.*

3 Learning Architecture for Human–Robot Natural Communication

We propose an architecture for learning natural human–robot joint attention. Overall architecture shown in Figure 2 has three features: (1) an autonomous attention selector that enables autonomous attention selection, (2) a self-organizing map to compress the high-dimensional face image into certain size, which makes learning time within a certain time period, and (3) a contingency learning module [11].

3.1 A learning process

The robot is programmed to gaze first at the caregiver's face, then at a salient object. Infants are supposed to have innate preference to the human faces

[12], and therefore, it may be natural to assume that the face is one of the most salient objects.

At the beginning of learning, the gate selects the output of visual feedback as the input to the robot. The robot will move its head, therefore, to gaze at the object that the module selects. When it succeeds to see the salient object in the center of the view, it will strengthen the connection between the face pattern and the joint displacements. Over time, the gate gradually use the output of the learning module more than the visual feedback. Note that the robot does not need any information whether joint attention successes or not, that is, the selected attention is not necessarily the same as that of the caregiver.

3.2 An automatic attention selector

In the previous work on acquiring human–robot joint attention, synchronization was pre-programmed [9, 11]. In the real communication, however, synchronization may not be pre-programmed but may emerge through interaction. Not only the robot but also the caregiver changes the behavior by seeing robot's behavior. Moreover, we do not know yet in what kind of preference to image features the robot should have to emerge the joint attention behavior. To study such resonance between the robot and the caregiver, the robot should autonomously change the gaze direction and see a different object according to its own interest measure changing from time to time.

A human, typically an infant, has a habituation ability, that is, he/she has a preference to a new and novel stimulus and to lose the interest when he/she attends to the same stimulus for a while. By losing the interest, a human changes the gaze direction. According to this observation, we propose an automatic attention selector to provide interest measure of each object that gradually decreases when the robot continues to see it. Note that this measure provides not only temporal change of the gaze, but also the preference change.

Typical image features such as color, edge, and motion are candidates to be attended to. Let n be the number of candidates in the robot's camera image and $S_i(t),\ (i = 1, \cdots, n)$ be the saliency of the i-th object, respectively. We set an initial value of the interest measure $I_i(t)$ as

$$I_i(0) = S_i(0)C_i, \tag{1}$$

where C_i is a weighting constant between different kinds of saliency, that is, a preference for the saliency. While the robot continues to gaze at the k-th object, the interest measure will decrease:

$$I_k(t+1) = \gamma_k I_k(t), \tag{2}$$

where $0 < \gamma_k < 1$ is a decay factor of the object k. The probability P_i to attend to the i-th object is calculated according to the interest measure,

$$P_i(t) = \frac{I_i(t)}{\sum_{j=1}^{n} I_j(t)}. \tag{3}$$

We can use more complicated functions such as softmax instead. On every time step, the robot will select an object k to attend to according to the probability (3). Over time, the robot will lose the interest measure according to the eq.(2), and eventually, it changes the object to attend to. We adopt visual feedback control for the robot to change the direction of gaze by moving pan and tilt angles.

3.3 A self-organizing map for categorizing face patterns

It is relatively hard to extract features concerning to the direction of the face since the face pattern has so much information. Therefore, if the face pattern is fed to the learning mechanism directly, the learning time will be large [9, 11].

We adopt a self-organizing map to compress the high-dimensional face image into a certain dimension, which makes the leaning time shorter. Before learning the contingency between the face pattern and the gaze direction, the robot sees the caregiver's face and categorizes its patterns by the self-organizing map.

3.4 A contingency learning module

The robot learns the contingency between the face pattern and the joint displacements [11]. We can use any simple learning network that can code one-to-one mapping such as a Hebbian network or a forward neural network for contingency learning.

4 Experiments

We realized the proposed architecture on a real robot agent and study on the learning performance of the joint attention behavior with different attending strategies. Here, several preliminary experimental results are shown.

4.1 An environmental setup

In Figure 3, we show an environmental setup for experiments. The robot and the human caregiver are sharing the same task space, a table in this example, where several objects are existing. The caregiver attends to objects by himself, and the robot will learn the contingency asynchronously. The caregiver changes the object positions randomly and asynchronously from time to time.

4.2 Self-organizing map for recognizing the face

A learned self-organizing map is shown in Figure 4. The size of self-organizing map is 9×9, each of which

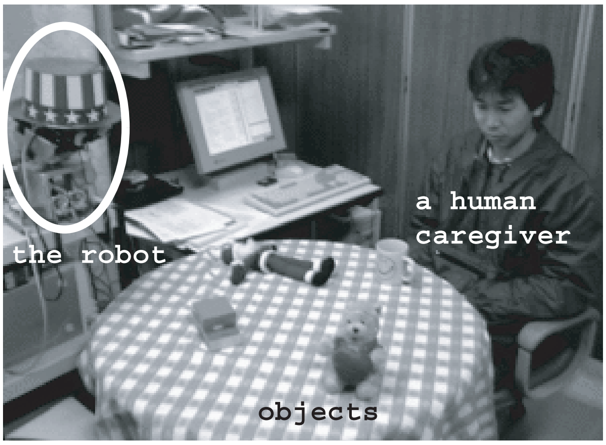

Figure 3: *An environmental setup for experiments: The robot and the human caregiver are seeing at objects on the table. The caregiver changes the object positions randomly and asynchronously from time to time.*

consists of 32 × 32 gray scale. This map can be learned within 3 [min].

In the previous work [11] in which they used 3-layer forward network, results of 125 trials are repeatedly fed to the network 5×10^5 times to learn the behavior, that is, each data is used 4000 times. By introducing the self-organizing map, we can adopt a 2-layer forward network instead and the learning time drastically decrease to several hundred trials without using data repeatedly. It needs 6 hours for 500 trials.

4.3 Attention selection between two objects

By introducing the automatic attention selector, we could reduce the learning time to 30 minutes for 500 learning epochs since the caregiver did not have to synchronize to the motion of the robot.

In Figure 5, we show how the robot changes its interest measure by the module. In this case, there are only two object in the environment. By image processing, intensity of each is obtained and used as saliency. Horizontal axis represents the frame number whose rate is 15 [Hz]. The module selects the attention every 20 frames, that is, every 1.32 [s].

When the object A was attended, the interest measure of A decreased while that of B did not change, and vise versa. At approx. the 700-th frame, we reset the measure of B since it became less than a given threshold ($\epsilon = 40$). It is the same reason for the sudden jump of the measure of A at approx. the 850-th frame.

By using this strategy, every object that provides more intensity than 40 can be attended, and over time, the interest measures for all objects become almost the same. Although it is controversial how to find appropriate strategy, the contingency learning module can make the robot learn the joint attention even with such a simple strategy.

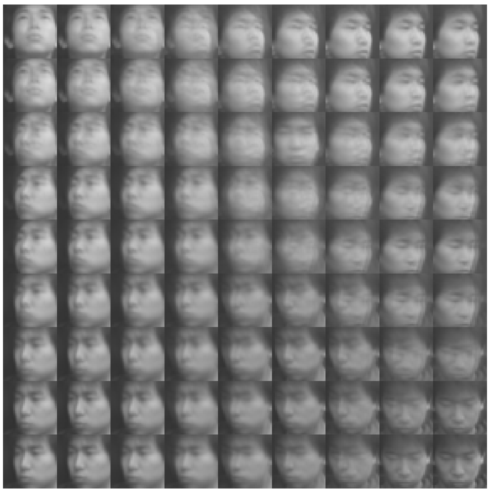

Figure 4: *An acquired self-organizing map: The size of self-organizing map is 9×9, each of which consists of a 32×32 gray scale bitmap. This map can be learned within* 3 *[min].*

4.4 Experiment 1: Changing attention selection speed

We changed the speed of attention selection both of the human and of the robot, and investigated the performance of learning. In Figure 6, we show moving average of success rate of joint attention in the last 50 trials, **case 1:** the caregiver slowly changes the attention trying to synchronize to the behavior of the robot, **case 2:** the attention selection speed of robot is faster than that of the caregiver, **case 3:** the attention selection speed of caregiver is faster than that of the robot, and **case 4:** the robot and the caregiver change the attention fast and asynchronously. In these experiments, there are only two objects in their view.

Note that the success rate at the 500-th step and that at the 50-th step mean different performances, that of the learning mechanism and that of the probabilistic gaze of the attention selector, respectively. We can see that synchrony let the robot increase the success rate from the chance level to almost 80% (case 1). Also in other cases, the robot still could acquire joint attention, which proves the inference that the contingency learning module can handle such asynchrony. It is astonishing that the proposed architecture can acquire joint attention even if the caregiver sometimes changes the gaze direction asynchronously while the robot gaze at the caregiver's face and at the object (case 4).

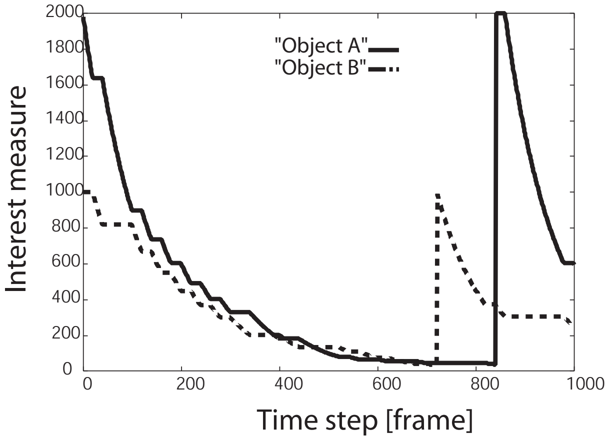

Figure 5: *Change of interest measures of object A and B: When the object A was attended, the interest measure of A decreased while that of B did not change, and vise versa. At approx. the* 700*-th frame, we reset the measure of B since it became less than a given threshold* ($\epsilon = 40$)*. It is the same reason for the sudden jump of the measure of A at approx. the* 850*-th frame.*

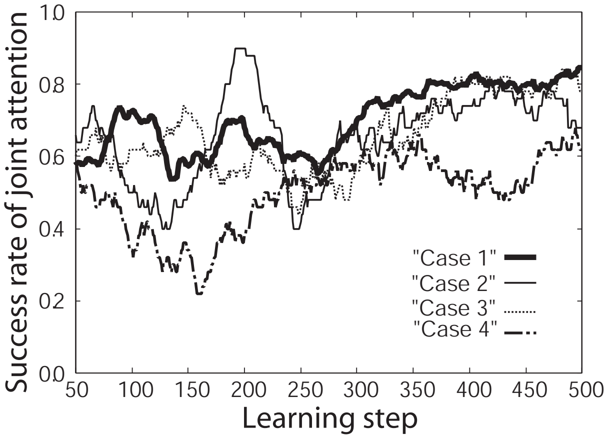

Figure 6: *Moving average (50 trials) of success rate changing the gaze: case 1: the caregiver slowly changes the attention trying to synchronize to the behavior of the robot, case 2: the attention selection speed of robot is faster than that of the caregiver, case 3: the attention selection speed of caregiver is faster than that of the robot, and case 4: the robot and the caregiver change the attention fast and asynchronously.*

4.5 Experiment 2: Changing preferences

Next, we investigated what does the difference of preference bring to the performance of the behavior. In this experiment, there are two objects, one is red and the other is yellow. Two cases are tested on the robot: (1) The robot has a preference to red whereas the caregiver has the same preference (see Figure 7) and (2) It has a preference to yellow whereas the caregiver has a preference to red (see Figure 8). The preference is realized by setting $I_{\mathrm{red}}(0) = 1000$, $I_{\mathrm{yellow}}(0) = 3000$, $\gamma_{\mathrm{red}} = 0.01$, and $\gamma_{\mathrm{yellow}} = 0.03$, respectively.

If the robot has the same preference with a caregiver, the success rate is relatively high from the beginning. However, even if the robot has a different preference, it can acquire the behavior. The robot even gaze at red since the probability to attend to red is not completely zero.

In the figures, we show two cases with and without decreasing the interesting measure according to eq.(3). We could not find any significant difference between them, which means that the contingency learning mechanism can handle both cases. However, it may depend on designing the gating function. If the gating is different, it is expected that the performances with and without the decay factor will differ. We have to continue to study on this matter.

5 Discussion

If we want to study the time development of the communication, the interaction dynamics consisting of dynamics of the caregiver (e.g. attending strategy), of the robot, and of the environment must be appropriately designed. The architecture was designed to realize such interaction dynamics by introducing an automatic attention selector, a self-organizing map, and a contingency learning module. To study further on the communication, we also have to measure the behavior of the caregiver, and to estimate his/her strategy to understand overall interaction dynamics. As for the dynamics of the environment, we can infer that as far as the environment is supposed to be quasi-static for the agents, that is, the objects do not move while the agents change the gaze, the proposed architecture can acquire joint attention.

The experimental results are still preliminary: we should do more experiments on more subjects to investigate in what kind of preference to image features the robot should have to emerge the joint attention behavior. The merit of taking such a constructivist approach is that we can estimate the validity of the attention mechanism in the context of learning joint attention, whereas it takes great pains to guess appropriate performance index to validate the mechanism itself.

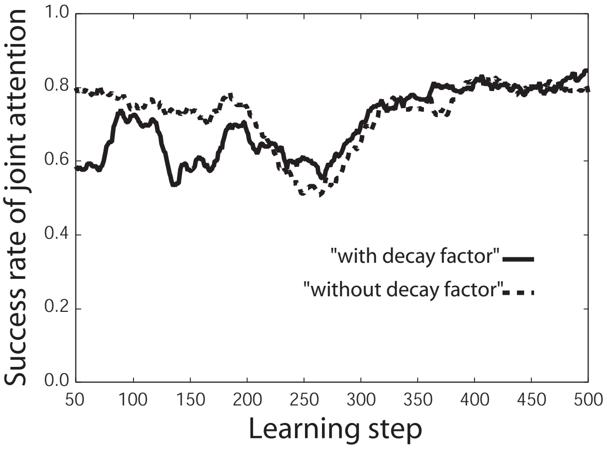

Figure 7: *Moving average (50 trials) of success rate (case 1: the same preference): Both the robot and the caregiver have the same preference to red.*

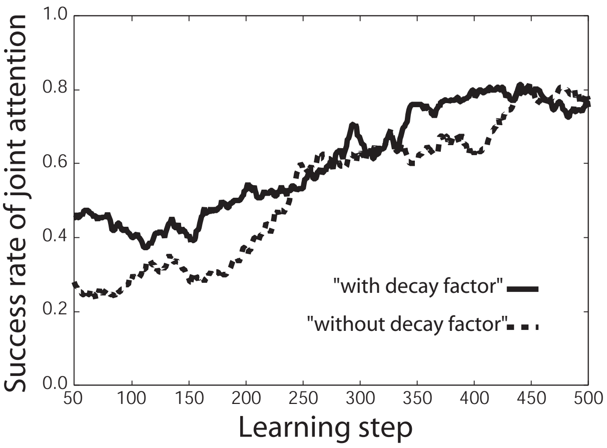

Figure 8: *Moving average (50 trials) of success rate (case 2: different preferences): The robot has a preference to yellow whereas the caregiver has a preference to red.*

It is also an interesting issue to study on the resonance between the robot and the caregiver. In this paper, the caregiver tried not to change the strategy while the robot autonomously change the gaze direction and see a different object according to its own interest measure. However, it must be interesting to study on the dynamical change of caregiver's strategy over time and on emergence of the synchrony.

We found that the short learning time is not only effective for preserving the caregiver's real-time change of dynamics, but also for statistical analysis. That is, if the leaning time is short, we can test several learning trials on several subjects easily, and as a result, we can process the results in a statistic way, which is necessary to analyze the communication between agents.

Acknowledgement

The authors wish to express our gratitude to Yukie Nagai and Yuichiro Yoshikawa for variable discussions. This study was performed through the Advanced and Innovational Research program in Life Sciences from the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

References

[1] G. E. Butterworth and N. L. M. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[2] M. Scaife and J. S. Bruner. The capacity for joint visual attention in the infant. *Nature*, 253:265–266, 1975.

[3] Chris Moore and Philip J. Dunham, editors. *Joint Attention: Its Origins and Role in Development.* Lawrence Erlbaum Associates, 1995.

[4] Simon Baron-Cohen. *Mindblindness.* MIT Press, 1995.

[5] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, 8(1):49–74, 2000.

[6] Hideki Kozima and Hiroyuki Yano. A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, 2001.

[7] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. In *Proceedings of 10th IEEE International Workshop on Robot and Human Communication*, 2001.

[8] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12:13–24, 2002.

[9] Yukie Nagai, Minoru Asada, and Koh Hosoda. Developmental learning model for joint attention. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 932–937, 2002.

[10] Ian Fasel, Gedeon O. Deák, Jochen Triesch, and Javier Movellan. Combining embodied models and empirical research for understanding the development of shared attention. In *Proceedings of the 2nd International Conference on Development and Learning*, pages 21–27, 2002.

[11] Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 12 2003.

[12] J. Gavin Bremner. *Infancy.* Blackwell, 1994.