Learning to Interpret Pointing Gestures: Experiments with Four-Legged Autonomous Robots

Verena V. Hafner and Frédéric Kaplan

Sony CSL Paris, 6 rue Amyot, 75005 Paris, France
{hafner, kaplan}@csl.sony.fr

Abstract. In order to bootstrap shared communication systems, robots must have a non-verbal way to influence the attention of one another. This chapter presents an experiment in which a robot learns to interpret pointing gestures of another robot. We show that simple feature-based neural learning techniques permit reliably to discriminate between left and right pointing gestures. This is a first step towards more complex attention coordination behaviour. We discuss the results of this experiment in relation to possible developmental scenarios about how children learn to interpret pointing gestures.

1 Introduction

Experiments with robots have successfully demonstrated that shared communication systems could be negotiated between autonomous embodied agents [1, 2, 3, 4, 5, 6]. In these experiments, robots draw attention through verbal means to an object of their environment. In order to bootstrap these conventional communication systems, it is crucial that the robots have a non-verbal way to influence the attention of other robots. They can for instance point to the topic of the interaction. This non-verbal form of communication is necessary as the robots have no direct access to the "meanings" used by the other robots. They must guess it using non-linguistic cues. The interpretation of pointing gestures must therefore be sufficiently reliable, at least initially when the system is bootstrapping. Once the language is in place, such kind of external feedback is less crucial and can even be absent [7].

Research in gaze or pointing interpretation is active in the context of human robot interaction (e.g. [8, 9, 10, 11, 12, 13]). By contrast, only few works explore the same issues for interaction between autonomous robots. A small number of solutions have been proposed to enable pointing and pointing interpretation in a variety of contexts (e.g. [14]). The focus of the present chapter concerns how robots can *learn* to interpret pointing gestures.

This chapter presents a model in which pointing gesture recognition is learned using a reward-based system. This model assumes, for instance, that a robot will often see something interesting from its point of view when looking in the direction where another robot is pointing to. It can be a particular salient feature of the environment, or an object which serves a current need (e.g. the charging station), or an opportunity for learning [15]. This approach is in line with Carlson and Triesch's computational model of the emergence of gaze following based on reinforcement learning [16]. Their model

S. Wermter et al. (Eds.): Biomimetic Neural Learning, LNAI 3575, pp. 225–234, 2005.

has been tested in a virtual environment by Jasso et al. [17]. To the best of our knowledge, this chapter represents the first attempt to show that a robot can learn to interpret the pointing gestures of another robot.

The rest of the paper describes the robotic experiment we conducted. We then discuss the limitation and possible extensions of this preliminary investigation.

2 Robot Experiments

2.1 The Interaction Scenario

Here we describe and show robot experiments where a pointing gesture is learned to be classified as either left or right. For these experiments, two Sony AIBOs were sitting on the floor, facing each other (see figure 1). One of the robots (the adult) is randomly pointing towards an object on the left or right side of its body using its left or right front leg, respectively. The other robot (the child) is watching it. From looking at the pointing gesture of the other robot, the learning robot guesses the direction and starts looking for an object on this side. Finding the object on this side represents a reward.

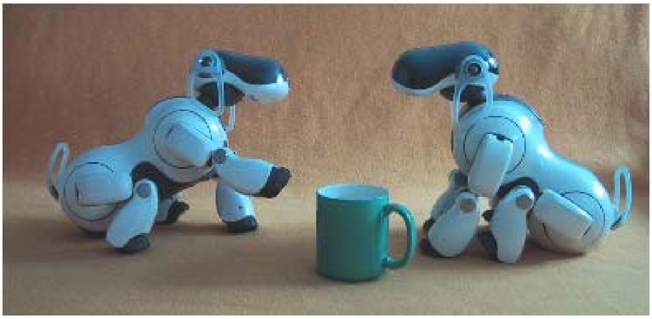

Fig. 1. An example of pointing shown with two robots. The robot on the left represents the adult who is pointing, the robot on the right represents the child who is learning to interpret the pointing gesture

Since the focus of this experiment is learning of pointing recognition and not pointing, this skill is hardwired in the adult robot. The robot is visually tracking a coloured object on its left or right side, thereby facing the object. Pointing is achieved by simply copying the joint angle of the head to the joint angle of the arm. Note that the pointing robot takes on an exact pointing position and does not only distinguish between the left and the right side.

2.2 Image Processing and Feature Space

A sample camera image from the robot's point of view can be seen in figure 2 left. For the experiments, the robot took 2300 pictures focusing on its pointing partner, 1150 for each pointing direction. The situations in which the pictures have been taken varied in the distance between the two robots, the viewing angle, the lighting conditions and the backgrounds (three different backgrounds).

From the original camera image, a small number of features has to be selected to facilitate the learning of interpreting the pointing gesture. We decided to apply two main filters to the image. One filter extracts the brightness of the image, the other filter extracts horizontal and vertical edges. These choices are biologically motivated. Eyes are very sensitive to brightness levels, and edges are the independent components of natural scenes [18]. The original image I is thus transformed to I' using a filter f:

$$I \xrightarrow{f} I'$$

For both filters, the colour image is transformed into greyscale first with pixel values between 0 and 255. In the subsequent steps, the image is divided into its left part and its right part (see figure 3). This is justified by the robot always centering on the other robot's face using an independent robot tracking mechanism, thus dividing the image into the right half of the other robot and its left half.

$$I' \longrightarrow I'_L, I'_R$$

The brightness filter B_θ applies a threshold θ to the image, which sets all pixels with a value greater than θ to 255, and all others to 0. For the experiments, values of $\theta = 120$ and $\theta = 200$ have been used. For the edge filter, we chose two Sobel filters S_H and S_V (see [19]) which extracts the horizontal and the vertical edges, respectively. An example of an image transformed by the filters can be seen in figure 2.

To the filtered images I', different operators op can be applied to extract low-dimensional features. These operators are the centre of mass $\mu = (\mu_x, \mu_y)$ and the sum Σ.

$$I' \xrightarrow{op} q$$

where q is the resulting scalar feature.

The four filters B_{120}, B_{200}, S_H and S_V together with the three operators μ_x, μ_y and Σ applied to both the left and the right side of the image I result in $4 \cdot 3 \cdot 2 = 24$

Fig. 2. Left: A robot pointing to its left side as seen from another robot's camera. The child robot tracks the adult robot in order to keep it in the centre of its visual field. Centre: Feature extraction for brightness using a threshold θ. Right: Feature extraction for horizontal edges using a Sobel edge detector

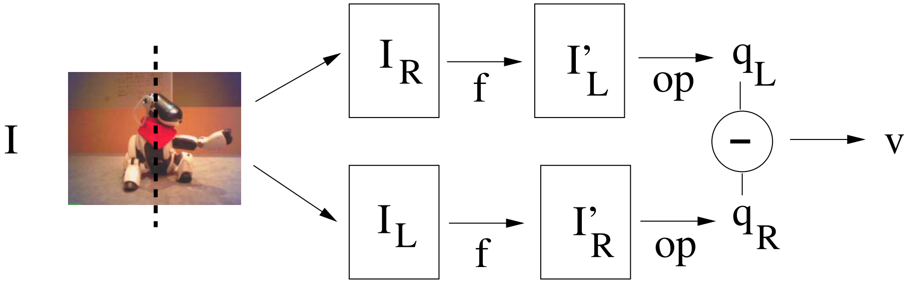

Fig. 3. Feature extraction from the original camera image

different features q_L and q_R (see figure 3). We take the differences between the left and right features resulting in 12 new features $v = q_L - q_R$.

2.3 Feature Selection

We selected a subset of the features by applying pruning methods. This is done by evaluating a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. The method used was greedy hillclimbing augmented with a backtracking facility provided by WEKA [20]. From the 12 features available to the robot, 3 have been selected to be the most meaningful: $B_{200} \circ \mu_y$, $S_H \circ \Sigma$ and $S_V \circ \Sigma$. Their values for all images are depicted in figure 4. Intuitively, the robot lifting its arm results in a vertical shift of brightness on this side of the image, an increase of horizontal edges and a decrease of vertical edges on this side.

For comparison, we also calculated the three least successful features. They turned out to be $B_{200} \circ \mu_x$, $B_{120} \circ \mu_x$ and $S_V \circ \mu_y$.

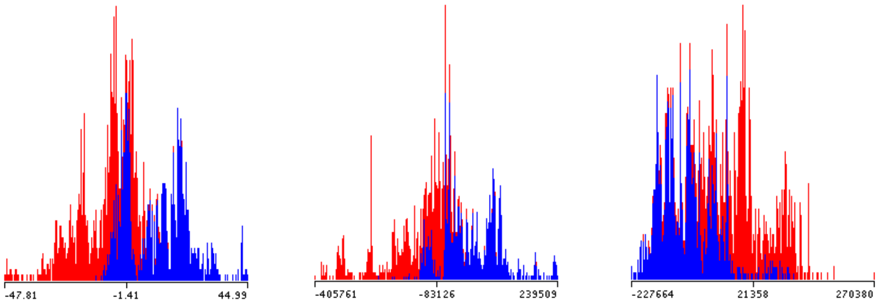

Fig. 4. Most successful scalar features for pointing gesture recognition from an image and the frequency of their values in the image data set. The red values are taken from pointing towards the left, the blue ones from pointing towards the right. Left: $B_{200} \circ \mu_y$. Centre: $S_H \circ \Sigma$. Right: $S_V \circ \Sigma$

3 Results

For learning the pointing gesture recognition, we used a multi-layer-perceptron (MLP) with the selected features as input, 3 neurons in the hidden layer, and the pointing direction (left or right) coded with two neurons as output. The learning algorithm is backpropagation with a learning rate $\lambda = 0.3$ and momentum $m = 0.2$. The evaluation is based on a 10-fold cross validation.

We chose backpropagation as a supervised learning algorithm which is comparable to a reward-based system in case of a binary decision. The choice of using MLPs and backpropagation is arbitrary and can be replaced by any other suitable machine learning technique involving reward. It is however sufficient to show that pointing gesture recognition can be easily learned between two robots.

Table 1. Learning results of different input features using 10-fold cross validation on the dataset of 2300 images

features	MLP	success rate
best 3	3-3-2	95.96%
worst 3	3-3-2	50.74%
all 12	12-7-2	98.83%

The success rate for the three chosen features (figure 4) is 95.96% (see table 1) using a 3-3-2 MLP and one epoch of training. When using all the 12 difference values v as inputs to a 12-7-2 MLP, the success rate increases to 98.83%. The success rate for the worst three features and one epoch of training is 50.74%, just slightly above chance.

In figure 5, the progress of learning can be monitored. The upper graph shows the error curve when the images of the pointing robot are presented in their natural order, alternating between left and right. The lower graph shows the error curve for images presented in a random order from a pre-recorded sequence. The error decreases more rapidly in the ordered sequence, but varies when conditions are changed.

4 Discussion

4.1 Pointing Interpretation and Intentional Understanding

We showed that with the current setup, a robot can learn to interpret another robot's pointing gesture. Although the pointing gesture of the adult robot can vary continuously depending on the position of the object, the interpretation of the pointing direction is either left or right. This corresponds to primary forms of attention detection as they can be observed in child development. Mutual gaze between an adult and a child, a special case of attentional behaviour, occurs first around the age of three months. At the age of about six months, infants are able to discriminate between a left or right position of the head and gaze of their parents, but the angle error can be as large as 60 degrees [21]. At the age of about nine months, the gaze angle can be detected correctly. Pointing gestures only start to be interpreted at the age of around one year [21] (see table 2). Children start

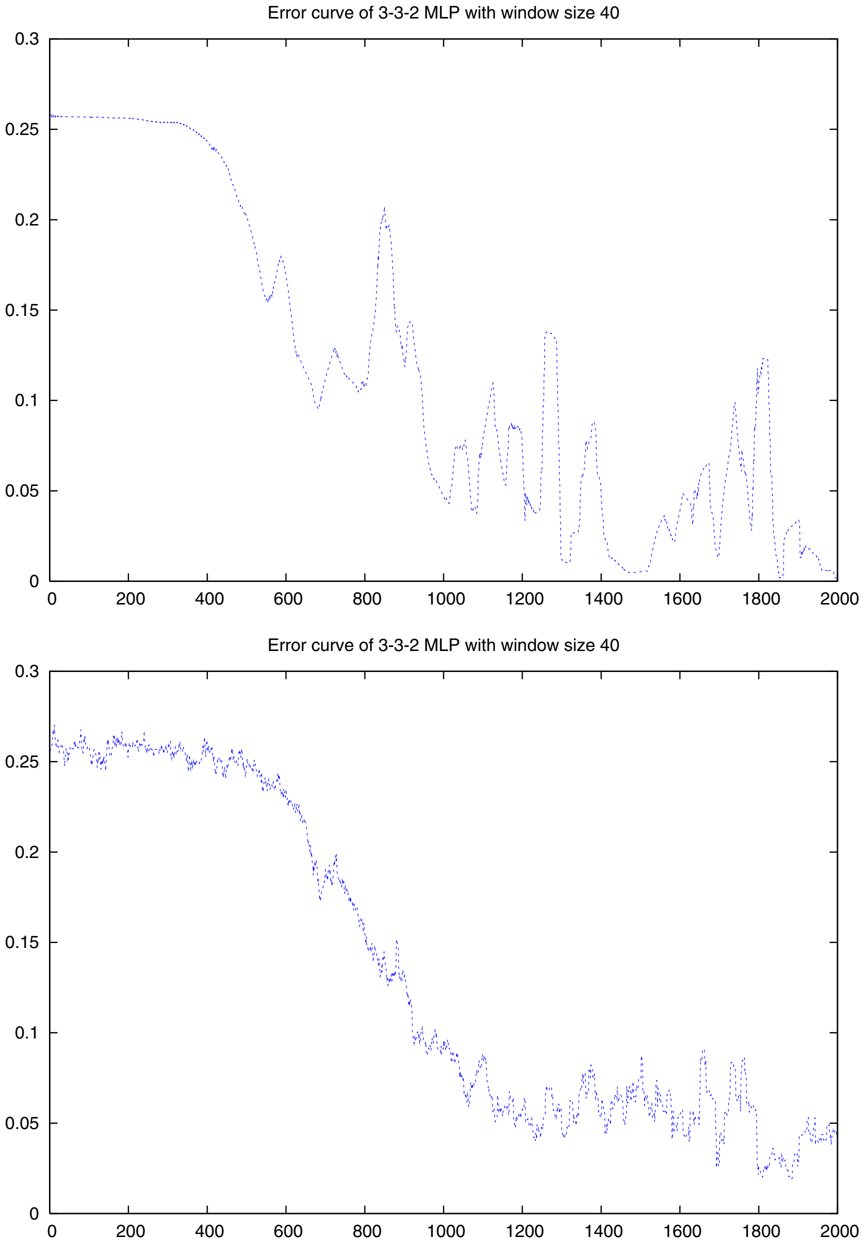

Fig. 5. Error of MLP during learning. Top: sequence of images in natural order. Bottom: random order of training images

to point first at the age of 9 months [22]. It is usually seen as a request for an object which is outside the reach of the child, and even occurs when no other person is in the room. This is called imperative pointing. At the age of 12 months, pointing behaviour

Table 2. Developmental timelines of attention detection and pointing in humans

Age from:	**Attention detection**	**Attention manipulation**
0-3 m	**Mutual gaze** - Eye contact detection	
6 m	Discrimination between **left and right position** of head and gaze	
9 m	**Gaze angle detection** - fixation on the first salient object encountered	**Imperative Pointing**: Drawing attention as a request for reaching an object (attention not monitored)
12 m	**Gaze angle detection** - fixation on any salient object encountered - Accuracy increased in the presence of a pointing gesture	**Declarative Pointing**: Drawing attention using gestures
18 m	**Gaze following** toward object outside the field of view - Full object permanence	

becomes declarative and is also used to draw attention to something interesting in the environment [23].

It is feasible to include the detection of a continuous angle of the pointing in a robotic setup. This would involve changing the current binary decision to a continuous value (possibly coded with population coding). But a higher accuracy is probably not sufficient to achieve efficient pointing interpretation. To truly learn the exact meaning of a pointing gesture, deeper issues are involved. Pointing interpretation in child development starts at an age where the infant begins to construct an intentional understanding of the behaviour of adults. This means that their actions are parsed as means towards particular goals. It could therefore be argued that pointing interpretation is much more than a geometrical analysis [23]. It involves a shared intentional relation to the world [24]. Developing some form of intentional understanding in a robot is one of the most challenging unsolved problems for developmental robotics [25].

4.2 Co-development of Pointing Gestures and Pointing Interpretation

In our robotic setup, the meaning of one gesture meaning ‘left’ and another gesture meaning ‘right’ could easily be reversed, or even completely different gestures could be used. The pointing movement of the adult robot was arbitrarily chosen to resemble a human pointing gesture. It is not clear that this gesture is the most adapted for unambiguous interpretations given the perceptual apparatus of the robots. In this perspective, it would be interesting to investigate a co-development between a pointing robot and a robot trying to understand pointing gestures. Situations of co-development between pointing and pointing gesture recognition could lead to interesting collective dynamics. Given the embodiment of the robots and the environmental conditions, some particular gestures may get selected for efficiency and learnability. Features that make them unambiguous and easy to transmit will be kept, whereas inefficient traits should be discarded. It has been argued that similar dynamics play a pivotal role for shaping linguistic systems [26].

4.3 Pointing and the Mirror Neuron System

Taking inspiration from current research in artificial mirror neuron systems [27], it would be possible to design a robot that interprets pointing gestures of others in relation with its own pointing ability. However, it is not clear whether the ability of pointing and pointing detection are correlated in human child development. Desrochers, Morisette and Ricard [28] observed that pointing seems to occur independently of pointing gesture recognition during infant development. These findings also seem to suggest that pointing does not simply arise from imitative behaviour.

4.4 Adult Robot Behaviour and Scaffolding

In the current setup, the adult robot randomly points at objects, its behaviour does not depend on the behaviour or the reaction of the child robot. Interactions between humans are very different. When pointing at something to show it to the child, a human adult carefully observes the attentional focus of the child and adjusts its behaviour to it. In some cases, the adult might even point to an object the child is already paying attention to in order to strengthen the relationship [29].

4.5 Pointing and Cross-Correlation

Nagai et al. [12] have argued in the context of human-robot interaction that simply the correlation between the presence of objects in general and gaze is sufficient for learning how to interpret gaze (without the necessity of an explicit feedback). Similar techniques based on cross-correlation could also be tried in the context of pointing interpretation between two robots. This type of learning relies on the assumption that the correlation is sufficiently strong to be discovered in practice. It is possible that a combination of both cross-correlation and reward based processes results in an efficient strategy for learning of pointing interpretation.

5 Conclusions

The interpretation of pointing is only one of the prerequisites necessary for bootstrapping human-like communication between autonomous robots. This chapter presents a first experiment showing how a robot can learn to interpret pointing gestures of another robot. In our future work, we will address the limitations of this initial prototype that have been discussed, and investigate the dynamics of social coordination and attention manipulation not yet investigated in this work.

Acknowledgements

This research has been partially supported by the ECAGENTS project founded by the Future and Emerging Technologies programme (IST-FET) of the European Community under EU R&D contract 001940. Andrew Whyte designed motor primitives for the robots used in this experiment. The authors would like to thank Anne Henning, Tricia Striano and Luc Steels for precious comments on the ideas discussed in this chapter.

References

1. Steels, L., Kaplan, F.: Situated grounded word semantics. In Dean, T., ed.: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence IJCAI'99, San Francisco, CA., Morgan Kaufmann Publishers (1999) 862–867
2. Steels, L., Kaplan, F.: Collective learning and semiotic dynamics. In Floreano, D., Nicoud, J.D., Mondada, F., eds.: Advances in Artificial Life (ECAL 99). Lecture Notes in Artificial Intelligence 1674, Berlin, Springer-Verlag (1999) 679–688
3. Steels, L., Kaplan, F.: Bootstrapping grounded word semantics. In Briscoe, T., ed.: Linguistic evolution through language acquisition: formal and computational models. Cambridge University Press, Cambridge (2002) 53–73
4. Steels, L.: The Talking Heads Experiment. Volume 1. Words and Meanings, Antwerpen (1999)
5. Vogt, P.: Lexicon grounding on mobile robots. PhD thesis, Vrije Universiteit Brussel (2000)
6. Kaplan, F.: La naissance d'une langue chez les robots. Hermes Science (2001)
7. Steels, L., Kaplan, F., McIntyre, A., Van Looveren, J.: Crucial factors in the origins of word-meaning. In Wray, A., ed.: The Transition to Language. Oxford University Press, Oxford, UK (2002) 252–271
8. Scassellati, B.: Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In: Computation for metaphors, analogy and agents. Vol 1562 of Springer Lecture Notes in Artificial Intelligence. Springer Verlag (1999)
9. Kozima, H., Yano, H.: A robot that learns to communicate with human caregivers. In: First International Workshop on Epigenetic Robotics (Lund, Sweden). (2001)
10. Imai, M., Ono, T., Ishiguro, H.: Physical relation and expression: Joint attention for human-robot interaction. In: Proceedings of the 10th IEEE International Workshop on Robot and Human Communication. (2001)
11. Nagai, Y., Asada, M., Hosoda, K.: A developmental approach accelerates learning of joint attention. In: Proceedings of the second international conference of development and learning. (2002)
12. Nagai, Y., Hosoda, K., Morita, A., Asada, M.: A constructive model for the development of joint attention. Connection Science **15** (2003) 211–229
13. Nickel, K., Stiefelhagen, R.: Real-time recognition of 3d-pointing gestures for human-machine-interaction. In: Proceedings of the 25th Pattern Recognition Symposium - DAGM'03. (2003)
14. Baillie, J.C.: Grounding symbols in perception with two interacting autonomous robots. In Berthouze, L., Kozima, H., Prince, C., Sandini, G., Stojanov, G., Metta, G., Balkenius, C., eds.: Proceedings of the 4th International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic System, Lund University Cognitive Studies 117 (2004)
15. Kaplan, F., Oudeyer, P.Y.: Maximizing learning progress: an internal reward system for development. In Iida, F., Pfeifer, R., Steels, L., Kuniyoshi, Y., eds.: Embodied Artificial Intelligence. LNAI 3139. Springer-Verlag (2004) 259–270
16. Carlson, E., Triesch, J.: A computational model of the emergence of gaze following. In: Proceedings of the 8th Neural Computation Workshop (NCPW8). (2003)
17. Jasso, H., Triesch, J., Teuscher, C.: Gaze following in the virtual living room. In Palm, G., Wermter, S., eds.: Proceedings of the KI2004 Workshop on Neurobotics. (2004)
18. Bell, A.J., Sejnowski, T.J.: Edges are the independent components of natural scenes. In: Advances in Neural Information Processing Systems (NIPS). (1996) 831–837
19. Dudek, G., Jenkin, M.: Computational principles of mobile robotics. Cambridge University Press (2000)
20. Witten, I., Eibe, F.: Data mining. Morgan Kaufmann Publishers (2000)

21. Butterworth, G.: Origins of mind in perception and action. In Moore, C., Dunham, P., eds.: Joint attention: its origins and role in development. Lawrence Erlbaum Associates (1995) 29–40
22. Baron-Cohen, S.: Mindblindness: an essay on autism and theory of mind. MIT Press, Boston, MA, USA (1997)
23. Tomasello, M.: Joint attention as social cognition. In Moore, C., Dunham, P., eds.: Joint attention: its origins and role in development. Lawrence Erlbaum Associates (1995) 103–130
24. Hobson, P.: The craddle of thought. MacMillan (2002)
25. Kaplan, F., Hafner, V.: The challenges of joint attention. In Berthouze, L., Kozima, H., Prince, C., Sandini, G., Stojanov, G., Metta, G., Balkenius, C., eds.: Proceedings of the 4th International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic System, Lund University Cognitive Studies 117 (2004) 67–74
26. Brighton, H., Kirby, S., Smith, K.: Cultural selection for learnability: Three hypotheses underlying the view that language adapts to be learnable. In Tallerman, M., ed.: Language Origins: Perspective on Evolution. Oxford University Press, Oxford (2005)
27. Elshaw, M., Weber, C., Zochios, A., Wermter, S.A.: A mirror neuron inspired hierarchical network for action selection. In Palm, G., Wermter, S., eds.: Proceedings of the KI2004 Workshop on NeuroBotics, Ulm, Germany (2004) 98–105
28. Desrochers, S., Morisette, P., Ricard, M.: Two perspectives on pointing in infancy. In Moore, C., Dunham, P., eds.: Joint Attention: its origins and role in development. Lawrence Erlbaum Associates (1995) 85–101
29. Liszkowski, U., Carpenter, M., Henning, A., Striano, T., Tomasello, M.: Twelve-month-olds point to share attention and interest. Developmental Science **7** (2004) 297 – 307