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

This paper explores the hypothesis that pointing gesture recognition can be learned using a reward based system. An experiment with two four-legged robots is presented. One of the robots takes the role of an adult and is pointing to an object, the other robot, the learner, has to interpret the pointing gesture correctly in order to find the object. We discuss the results of this experiment in relation to possible developmental scenarios about how children learn to interpret pointing gestures.

1 Introduction

The skill of pointing interpretation is a crucial step in the development of a young child toward the capability of joint attention [1]. It plays an important role in the non-linguistic communication system which serves as a foundation for the development of language [2]. The comprehension of pointing seems to require more than gaze detection, and might be connected with the apprehension of complex signs between 10 and 12 months [3]. How does this capability develop?

This paper explores the hypothesis that pointing gesture recognition is learned using a reward-based system. This hypothesis assumes, for instance, that by looking in the direction where the adult is pointing to, the child will often see something interesting from its point of view. It can be an interest in an opportunity for learning [4] or in an object which serves a current need (e.g. food). It can happen that by misinterpreting the pointing direction the child will also find an interesting object, however, the hypothesis assumes that there is a statistically significant correlation between the correct pointing direction and interesting objects. Alternatively, the hypothesis concerns situations where parents reward the child with affection or another emotional response if the child interprets the adult's pointing gesture correctly. In such a case, the child would learn the pointing interpretation through a sort of parental conditioning.

The learning process in both situations shares a common underlying structure. It can be modelled as a supervised learning process with the pointing gesture as the visual input, and the location of the pointing aim, or the movement direction to look at this location, as the desired output. The evaluation feedback is obtained through the reward, which is the satisfaction of a drive in the first case and an emotional response in the parental conditioning case.

In the rest of the paper, we will show that a robot can learn to interpret pointing gestures of another robot using a reward-based system. We will then discuss the results of this experiment in relation to potential underlying mechanisms described in child development literature.

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.

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 focussing 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

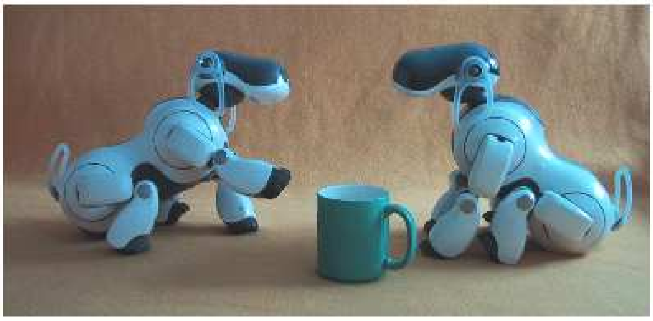

Figure 1: An example of pointing shown with two robots. The robot on the left represents the adult who is pointing, the robot on the right represents the child who is learning to interpret the pointing gesture.

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 [5]. 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 [6]) 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.

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

Figure 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.

where q is the resulting scalar feature.

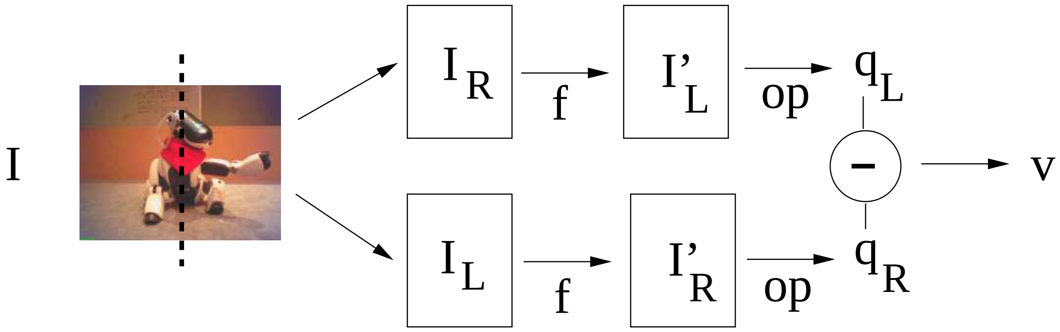

Figure 3: Feature extraction from the original camera image.

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 $4 \cdot 3 \cdot 2 = 24$ 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 [7].

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.

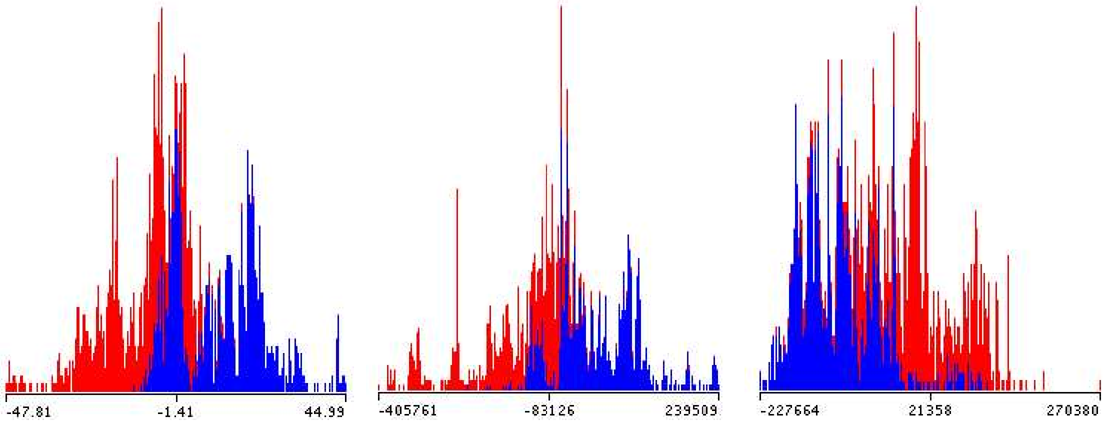

Figure 4: Most successful scalar features for pointing gesture recognition from an image and the frequency of their values in the image data set. The red values are taken from pointing towards the left, the blue ones from pointing towards the right. Left: $B_{200} \circ \mu_y$. Centre: $S_H \circ \Sigma$.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

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 $S_V \circ \mu_y$.

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 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 3) 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

Using a setup with two robots, we showed that pointing gesture detection can be easily learned with a reward based system. To the best of our knowledge, this is the first experiment on pointing and pointing gesture recognition using two robots. Earlier robotic experiments focused on gaze and pointing gesture recognition between a human and a robot [8, 9]. A computational model of reward-based emergence of gaze following was presented in Carlson and Triesch [10]. Their findings which can probably be extended to pointing interpretation are compatible with ours. But there are still other hypotheses about the origin of pointing gesture detection.

By the age of 9 months, the child becomes capable of imperative pointing [11]. Pointing can for instance be used to ask for an unreachable object, but at this stage, the child does not monitor the attention of the adult. Could it be possible that the child interprets the pointing gestures of others in relation with its own pointing ability? Are mirror neurons playing a role in this interpretation [12]? This remains an unlikely hypothesis, since no experimental correlation has been found in children between the learning of pointing and the learning of pointing gesture recognition [13].

Another possibility could be that simply the correlation between the presence of objects in general and pointing gestures is sufficient for learning how to interpret pointing

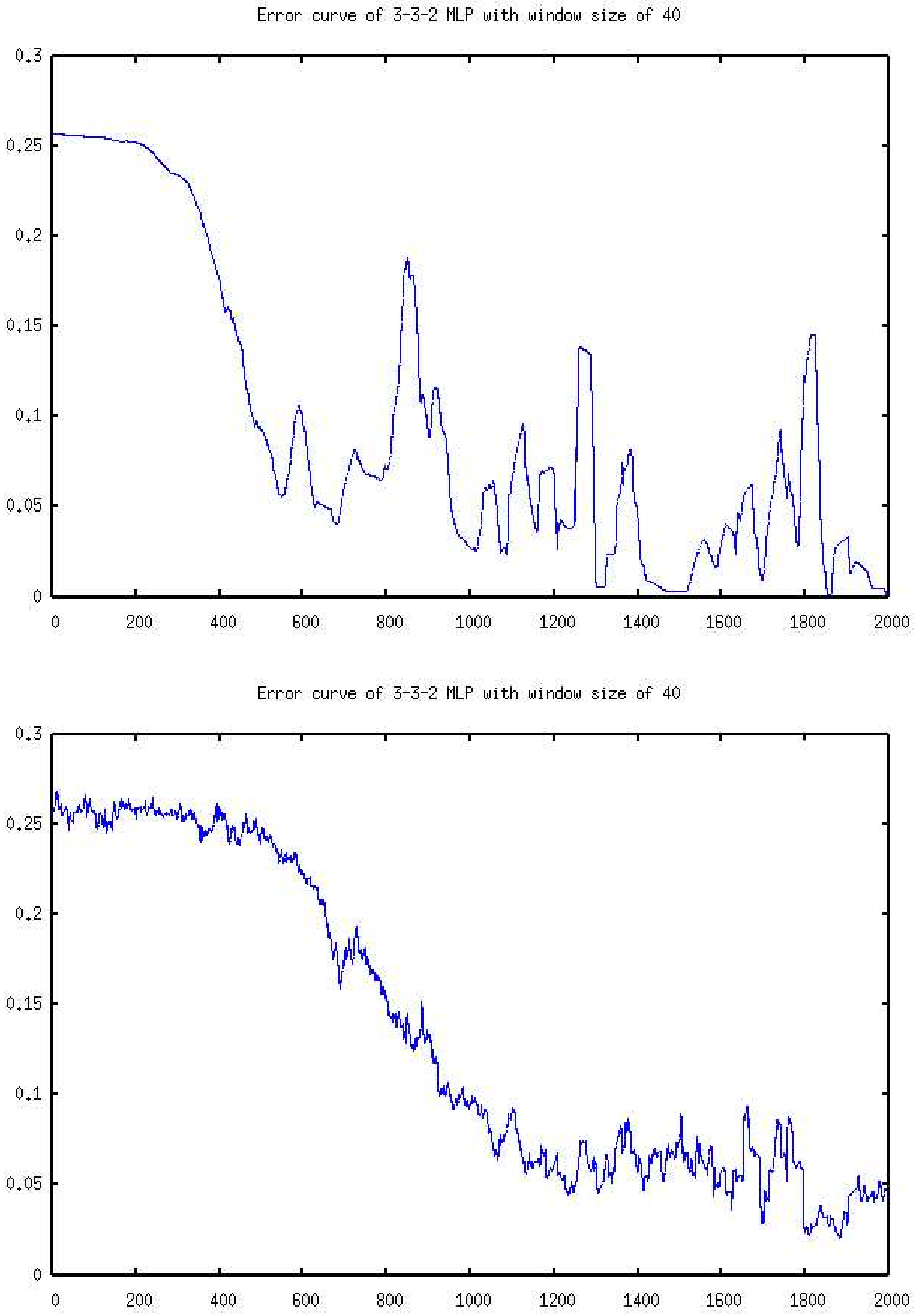

Figure 5: Error of MLP during learning. Top: sequence of images in natural order. Bottom: random order of training images.

(without the necessity of an explicit feedback). It has been tested in the case of gaze following in the context of a human-robot interaction by Nagai et al. [8] and found to be feasible. This hypothesis relies on the assumption that the correlation is sufficiently strong to be discovered in practice.

There is no definitive way of comparing the relative plausibility of the reward based and the correlation based hypothesis. It is possible that a combination of these processes is involved in the learning of pointing interpretation.

5 Perspectives

The interpretation of pointing is only a small piece of the developmental puzzle that leads to the capability of joint attention. In particular, we still need to understand the dynamics of social coordination and of attention manipulation. More importantly, to reach joint attention, a robot must monitor and direct the intention underlying the behaviour of others [1]. More experiments will be needed to progressively understand the development of these skills.

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.

References

[1] Kaplan, F., Hafner, V.: The challenges of joint attention. In: Proceedings of the 4th International Workshop on Epigenetic Robotics. (2004)

[2] 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

[3] Piaget, J.: The origins of intelligence in children. Norton, New York (1952)

[4] 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. Springer-Verlag (2004)

[5] 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

[6] Dudek, G., Jenkin, M.: Computational principles of mobile robotics. Cambridge University Press (2000)

[7] Witten, I., Eibe, F.: Data mining. Morgan Kaufmann Publishers (2000)

[8] Nagai, Y., Hosoda, K., Morita, A., Asada, M.: A constructive model for the development of joint attention. Connection Science **15** (2003) 211–229

[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] Carlson, E., Triesch, J.: A computational model of the emergence of gaze following. In: Proceedings of the 8th Neural Computation Workshop (NCPW8). (2003)

[11] Baron-Cohen, S.: Mindblindness: an essay on autism and theory of mind. MIT Press, Boston, MA, USA (1997)

[12] Gallese, V., Fadiga, L., Fogassi, L., Rizolatti, G.: Action recognition in the premotor cortex. Brain **119** (1996) 593–609

[13] 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