Motion-Based Robotic Self-Recognition

Philipp Michel
Department of Computer Science
Yale University
New Haven, CT, USA
Email: philipp.michel@yale.edu

Kevin Gold
Department of Computer Science
Yale University
New Haven, CT, USA
Email: kevin.gold@yale.edu

Brian Scassellati
Department of Computer Science
Yale University
New Haven, CT, USA
Email: scaz@cs.yale.edu

***Abstract*— We present a method for allowing a humanoid robot to recognize its own motion in its visual field, thus enabling it to distinguish itself from other agents in the vicinity. Our approach consists of learning a characteristic time window between the initiation of motor movement and the perception of arm motions. The method has been implemented and evaluated on an infant humanoid platform. Our results demonstrate the effectiveness of using the delayed temporal contingency in the action-perception loop as a basis for simple self-other discrimination. We conclude by suggesting potential applications in social robotics and in generating forward models of motion.**

I. Introduction

When deciding whether a primate is self-aware, evolutionary psychologists employ a controversial experiment called the mirror test [4]. When confronted with a mirror, a rhesus or stumptailed monkey will treat its reflection as another monkey, displaying signs of aggression. Chimpanzees, on the other hand, will quickly begin to use the mirror to preen themselves and pick at their teeth. They realize that their mirror reflection is associated with their own physical self and begin to exhibit self-directed behavior. While the chimpanzees learn quickly, the other species of monkey never become aware of this correspondence. Thus, the chimpanzees are considered self-aware, but the other monkeys are not. The mirror test not only plays a crucial role in the study of animal behavior [5], it also reveals insight into the development of self-awareness in humans. When confronted with an image of its body that is temporally contingent with its movement, such as a mirror reflection, a human infant exhibits signs of self-exploratory behavior from 3 months of age onwards. As early as at four months of age, babies display some measure of social self-awareness, showing more interest in experimenters imitating them than in their own reflections [7].

Most would agree that it is premature to label a robot that can identify its reflection in a mirror self-aware. We would probably expect self-aware robotic agents to possess introspection and reflection abilities leading to a far more complex sense of self than that afforded by simple visual self-identification. However, even if passing the mirror test is not a sufficient condition for self-awareness, it is still useful for a humanoid robot to be able to identify itself under a wide variety of circumstances [3]. Identification of which objects in the visual field are 'self' can serve as a foundation for a more complicated kinematic model, with expected trajectories for the results of self-movement. Furthermore, self-recognition provides a framework for grounding language concepts such as 'I' and 'myself' in perceptual experience. In social robotics, distinguishing self from other can aid in identifying social cues [2] and in mapping another agent's actions onto the self, thereby establishing a joint meaning of observed agent behavior. It has been argued that a sensory-motor approach to self-recognition may constitute a promising step towards basic self-awareness in robotic systems [1]. At the same time, it can inform models of self-awareness in evolutionary and developmental psychology, perhaps reopening debate on the significance of the mirror test.

Given these potential applications, there has been surprisingly little experimental work done on the subject of robotic self-recognition. In [1], the authors went only so far as to suggest simulating a robot that could recognize itself in a mirror, explaining that actually implementing such a system would be too difficult. Moreover, they did not suggest a specific mechanism for how the robot might begin to recognize itself. While work is underway at the University of Minnesota on building a robot that can identify scene motion caused by its own camera pan and tilt [6], it is not clear how their method will generalize to motion that is not over the robot's entire visual field. Motion caused by camera tilt has a much more predictable effect on the visual input than the motion of an arm, as the motor encodings at each joint do not correlate well with any specific properties of the image.

We have implemented a simple method for self-recognition on Nico, an infant-like humanoid robot currently in development at Yale. Nico learns through experimentation to expect motion in its visual field within a certain time window after initiating an arm motor movement. Once a representation of this characteristic time delay is present, motion regions in the visual field that appear within the learned time frame are labeled as 'self', with that label persisting for regions of motion in subsequent frames that are sufficiently similar. Note that this method intentionally avoids using kinematic models. This allows the robot to recognize itself under a wide variety of transformations to its physical structure or appearance in the scene. It also allows the robot to recognize its own motion in a mirror.

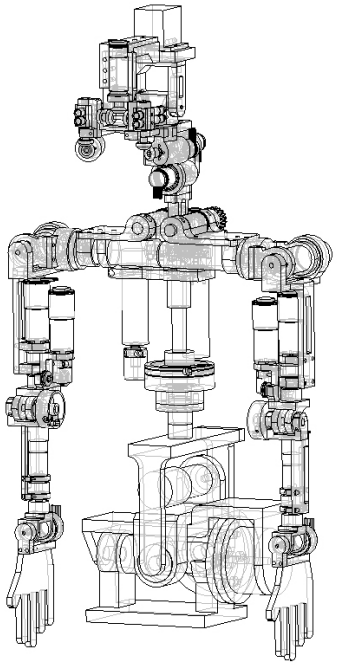


Fig. 1. Line drawing of the robot's current mechanical design.

II. Methodology

A. The Robot Platform

Our robot, Nico, is an upper-torso humanoid designed to resemble a one-year-old infant in both physical appearance and cognitive abilities. Still in development, it will serve as a robotic test-bed for theories of human social learning. Fig. 1 shows an outline of the current physical design.

Nico's active vision head accomodates two miniature CCD cameras for each eye, providing both wide and narrow fields of view, thus approximating foveate stereo vision in humans. For the evaluation purposes of this paper, we used the wide field of view cameras, although our approach is independent of the particular camera or lens characeristics. Overall, the head-neck assembly (shown in Fig. 2) has seven degrees of freedom (DOFs). Both eyes are equally affected by all head and neck movement, except for an additional degree of yaw that can be independently specified for each eye, implementing eye vergence.

Nico's six DOF arm is driven by miniature DC motors and can be maneuvered through the entirety of the robot's field of view and beyond. For our experiments, all arm joint movement was constrained to a set of angles that forced the arm to remain in the field of view at all times.

All vision processing and motor control is accomplished by a cluster of 16 processors running the QNX Neutrino RTOS connected by a 100Mbit switch. Communication and data transfer between nodes proceeds through a port-based interface, essentially implementing concurrency-safe shared memory between processors. Four frame grabbers acquire 320×240 pixel frames at 30Hz from the cameras. Subsequent vision processing takes place at 15Hz.

Fig. 2. The robot's head-neck assembly, housing a four camera vision system and providing a total of 7 DOFs.

B. Vision & Attention Processing

Scene data captured by the cameras passes several stages of visual and attentive processing before it can act as input to the motion delay learning module. Fig. 3 gives an overview.

First, the intrinsic and extrinsic camera parameters are used to undistort the camera image to yield a straight view of the scene. The calibration process needs to be executed only once after the cameras are fixed to their mounts and involves moving and tilting a checkerboard pattern around in front of the robot. Afterwards, a look-up table is used to undistort the incoming video stream on-the-fly.

A motion module performs image differencing on subsequent frames of the undistorted image stream to determine areas of motion. Incoming images are stored in a ring of three buffers: one for the current image I_0, one for the previous image I_1, and one for receiving new input. The module calculates a thresholded absolute value of the difference between the grayscale values in each image ($I_{raw} = \mathcal{T}(|I_0 - I_1|)$). It thus computes a raw monochromatic motion saliency map, with brighter pixels corresponding to more perceived motion.

The saliency map is passed to a module implementing a model of pre-attentive vision (PAV) in humans. It identifies regions of interest from saliency maps computed by a range of vision processors including color, face, skin and motion detectors. PAV computes an overall saliency map from the weighted sum of the individual maps, with weights being determined by the robot's current attentive configuration. In our experiments, the motion module was the sole contributor to the final saliency map. PAV tags the pixels of each individual region of interest with a unique identifier and places them within a bounding box. This process is repeated for each frame.

The final stage of processing consists of a memory module implementing simple object permanence. It associates bounded regions of motion across subsequent frames by comparing their shape and location. If two regions are sufficiently similar, they are considered as corresponding to the same moving object and given the same object identifier.

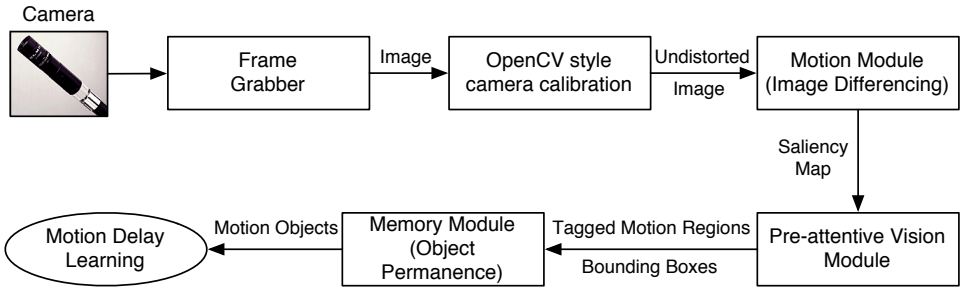


Fig. 3. Visual and attentive processing preceding the delay learning stage. A separate processing flow is associated with each eye.

The ultimate output of vision and attention processing thus consists of a set of moving objects, defined by bounding boxes with associated information such as extents and centroid. Each possesses a numerical identifier that can be used to keep track of the object as the motion proceeds.

C. Self-Recognition

Our implementation of self-recognition conceptually consists of two components. The core module incrementally learns the characteristic time delay inherent in the action-perception loop from a sequence of random arm motions within the visual field. A separate classification module uses the learned delay model to identify newly occurring moving objects that satisfy the delay window (thus conceptually belonging to the self) and highlights their salient pixels in the video stream.

To learn the characteristic time delay, a set of random arm poses is first assembled, constrained in a way such that all of them lie in the robot's field of view. Motor commands for each joint are then generated that move the arm through the sequence of random poses. Just before each set of commands is sent to the motors, we take a timestamp using QNX's real-time clock, which provides close to nanosecond accuracy. As the arm moves, we wait for the first time that a moving object is detected by the processing stage described above, taking another timestamp as soon as this happens. The temporal difference between those timestamps, t_1, is our current estimate of the time delay in the action-perception loop. Note that we only rely on generated poses for experimentation purposes. Motor commands may as well have been generated by any other program currently controlling the robot. The delay learning algorithm then simply bases its measurements on those movements.

In a similar manner, we measure the delay between physical completion of an arm movement and the time when no more moving objects are registered by the processing stage. During movement, the current position of every arm motor is compared to the desired target position for the current pose. Once the final arm position is reached, we take a timestamp and wait for the processing stream to assert that no more motion is present, taking another timestamp. The difference between the two timestamps, t_2, provides another delay measure of the action-perception loop. A timeline for the measurement process is given in Fig. 4.

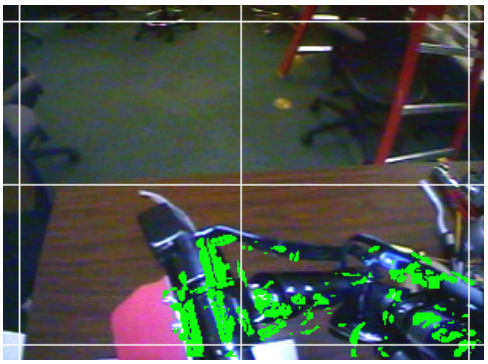


Fig. 5. Output from the self-motion classifier, overlaid onto the visual input from one eye. All salient pixels from a moving object identified as 'self' are highlighted (colored bright green).

The sequence of delay measurements for t_1 and t_2 allows us to iteratively refine the bounds on a characteristic time window within which we expect to visually perceive motion after having started an arm movement. These bounds, $[t_{1_{min}}, t_{1_{max}}]$ and $[t_{2_{min}}, t_{2_{max}}]$, are initialized to define an overly restrictive time window which is then gradually expanded to accomodate training data as it becomes available.

Our self-motion classifier takes as input the bounds $t_{1_{min}}$ and $t_{1_{max}}$ on the characteristic delay output by the learning module together with the start time of the last movement. It labels moving objects that first occur within the time window as belonging to the self. The object permanence implemented by the memory module then allows us to keep labeling these objects over the lifetime of the movement, resulting in highlighted (bright green) regions being tracked through the video stream. These correspond to the parts of the visual field recognized as 'self'. Our current classifier principally makes use of the bounds on t_1, disregarding $t_{2_{min}}$ and $t_{2_{max}}$, the bounds on the delay between the arm reaching its final position and the motion subsiding in the visual field. Preliminary results show that incorporating the bounds on t_2 can serve to reduce the rate of false positives. We expect to use those measurements in future work to provide a posteriori reassurance that a movement was correctly labeled. The result of a classification is shown in Fig. 5.

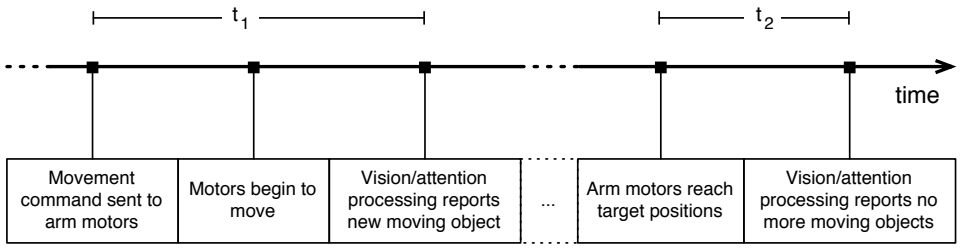


Fig. 4. Timeline showing relevant events for the measurement of t_1 and t_2.

III. Evaluation

A. Learning Rate

We first aimed to gain a quantitative understanding of the characteristic delay in the action-perception loop and how it varies over time as training data is acquired. Training proceeded under ideal conditions, with the robot arm constituting the sole source of motion in the scene. In more general scenarios, the learning module disregards delay measurements for which there is conflicting motion present in the scene at the time the motor commands are issued. To relax this condition, a simple outlier rejection mechanism can be used, allowing the learning module to handle a number of skewed measurements.

Fig. 6(a) shows that measured time delays for t_1 fluctuate about a mean of close to 500ms, a reasonable amount of time given the significant preprocessing taking place and the hardware involved. The fluctuations can be explained by the fact that certain arm movements cause more prominent motion than others, which is detected more quickly. Changes in processing load on the system over time also account for some of the variability.

Fig. 6(b) shows how the learned bounds on the characteristic time delay evolve as training data is acquired. The time window defined by the bounds gradually expands, changing only minimally after around 20 delay measurements. We found that after approximately 2 minutes of training, further changes in the learned delay bounds were negligible, hardly improving recognition accuracy.

B. Recognition Accuracy

The best-case recognition accuracy for different amounts of available training data was established next by determining the percentage of previously unseen arm movements correctly labeled as self-motion by the classifier module. During classification, the robot arm was again the sole source of motion in the visual field.

For a training set of 5 examples, 16 out of 50 movements were correctly classified, an accuracy of 32%. Given 10 training examples, 34 out of 50 movements, or 68%, were identified as self-motion correctly. Finally, for 25 training examples, the robot correctly labeled all 50 arm movements as self-motion.

Under these ideal conditions, a learning period of just over 2 minutes suffices to achieve very high recognition accuracy.

C. Self-Other Discrimination

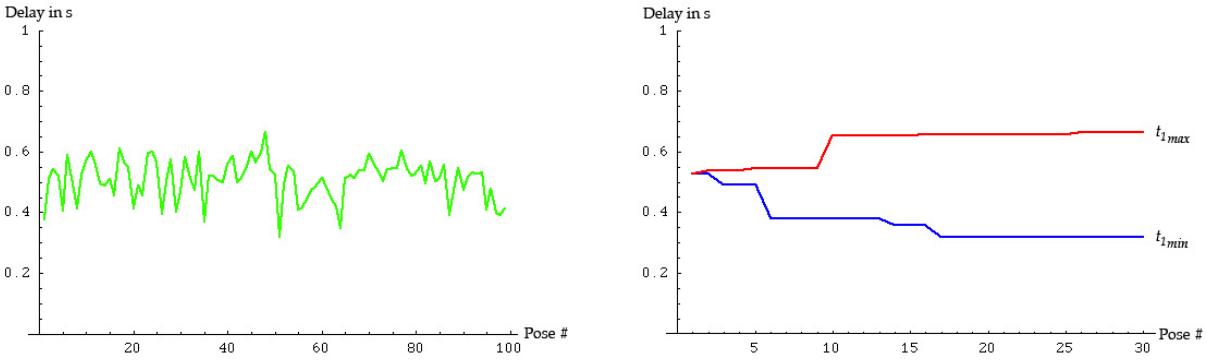
To evaluate recognition accuracy in the presence of human-induced motion in the visual scene, one of the authors moved his hand within Nico's field of view as soon as the robot successfully labeled an arm movement as self-motion. Essentially, the hand acted as a distractor for the object permanence module, which might group the human motion together with the robot motion and classify both as 'self'. Out of 75 random movements with the distractor, the hand's motion was mislabeled as self-motion 17 times, yielding 77.3% accuracy. Fig. 7 shows one trial.

Recognition accuracy dropped significantly when the distractor began to move in anticipation of the robot's motion. Out of 60 examples, the distractor's motion was falsely labeled as 'self' 33 times, yielding a false positive rate of 55%. Using the bounds on t_2 as an additional check, we were able to reduce the false positive rate to 20%. However, using both t_1 and t_2 also had the adverse effect of reducing Nico's self-recognition rate to a mere 66%. We are currently working on determining whether this tendency to mislabel nearly simultaneous movement is caused by excessive variability in the robot's mechanical response times, or if it instead implies that additional kinematic information is necessary to avoid mislabeling a human that is interacting with the robot. The problem might also be alleviated with a self-label that is attributed with confidence over time, rather than basing the decision solely on immediate input from pre-attentive vision.

D. Shape Independence / The Mirror Test

To demonstrate the benefits of using motion time delay as the sole basis for a self-recognition algorithm, we drastically changed the shape of the robot arm by covering it with a glove during recognition trials as shown in Fig. 8, having previously trained on an uncovered arm. Even though glove motion is actually less visually salient than motion caused by the arm alone, the recognition performance was not affected, as all 50 ungloved movements were correctly labeled as self-motion.

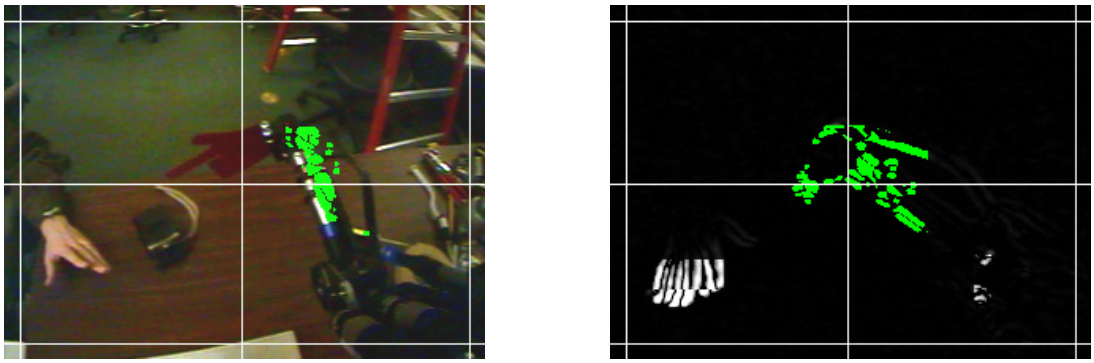
Finally, we attempted a rudimentary mirror-test on Nico, placing the robot about 2 feet from a large mobile mirror. At this distance, the classifier treated motion caused by the robot's body and its reflection as equivalent, successfully labeling the mirror reflection as 'self' whenever the arm movement satisfied the learned delay, as seen in Fig. 9. As the distance between robot and mirror is increased, accuracy gradually drops due to the decreasing area of



(a) Delay t_1 measured over a trial run of 100 arm movements. $\mu = 0.506599s$, $\sigma = 0.0660767s$.

(b) Bounds on the delay over a trial run of 30 arm movements.

Fig. 6. Measurements illustrating the learned characteristic time delay.



(a) First person view of the test condition with the distractor. Only the robot's motion is labeled as 'self'.

(b) Motion module output under the same conditions. Both the human hand and the robot arm are moving, but only the robot's motion satisfies the learned time delay (robot arm highlighted green, hand remains white).

Fig. 7. Simple self-other discrimination. A human distractor attempts to cause the classifier to falsely mark his motion as resulting from the robot's arm movement.

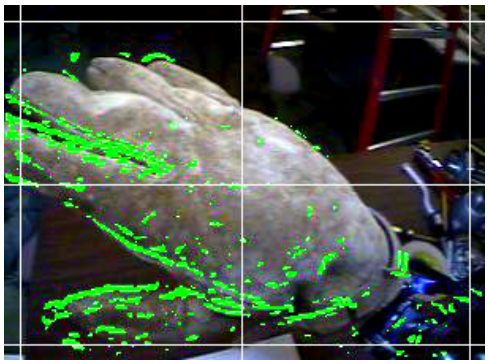


Fig. 8. Nico's gloved hand correctly being labeled as 'self'.

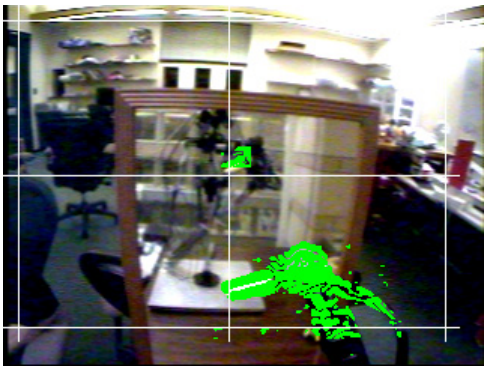


Fig. 9. Nico recognizes self-motion in a mirror.

pixel disparity between subsequent frames that the motion module can detect.

IV. Conclusion

Our results suggest that using a learned time delay is a promising method for identifying extensions of the self in the visual field. It has the advantages of versatility and

conceptual simplicity, extending naturally to identifying reflections as well.

As a model for the self-awareness displayed in human infants, the learned time delay model is necessarily incomplete. Infants have been shown to also expect a direction for their self-motion, and they are surprised if their reflections are reversed [8]–[10]. We intend to augment our approach with a mechanism for learning exactly such an expectation, using the detection of self-motion as a primitive in learning a more complex forward model. Again, such a method would identify reflected self-motion as well, since a reflected arm's overall motion vector roughly coincides with that of the physical arm. To the extent that such a forward model succeeded, the robot would likely be able to display the same qualitative performance in mirror tests as a 4-month-old infant.

We expect that a robust method of visually recognizing the robot's own physical presence will play a significant role in providing the humanoid with a perceptually grounded meaning of linguistic concepts such as 'I', 'myself' or, inversely, perhaps 'you, the robot' and 'Nico'. These form a crucial part in describing manipulation tasks a human might want the robot to learn from interaction, which constitutes a future area of our research.

One final application of learned time delays could be in identifying other social agents that are interacting with the robot. By associating a second characteristic time window with humans' reactions to his movements, for example during an imitation-based interaction, Nico could distinguish between individuals in the room who are actively engaging him socially and those who are not. Such information would be useful in directing attention in social situations, and might serve as a primitive in learning the social concepts of 'self' and 'other'. Furthermore, the ability to recognize socially responsive agents might allow the robot to attribute intents, beliefs and goals to the agent's actions, thus providing a first crucial step towards a robotic theory of mind.

ACKNOWLEDGMENTS

Support for this work was provided by a National Science Foundation CAREER award (#0238334). Some parts of the architecture used in this work was constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots) and from the DARPA CALO project (BAA 02-21). Additional support for one of the authors was provided by the German National Merit Foundation.

REFERENCES

[1] L. Berthouze and S. Itakura. "Possibility of Self-Recognizing Robots: From the Perspective of Research on Nonhuman Primates." *Japanese Journal of Cognitive Studies: Consciousness: Toward a Cognitive Science of Brain and Mind (Special Issue)*, Vol. 4(3), pp. 120–127, 1997.

[2] C. Breazeal, K. Dautenhahn, and B. Scassellati. "Recognition of self and other through imitation games." http://www.cs.yale.edu/homes/scaz/abstracts/1999/scaz2.pdf

[3] H. Cruse. "The evolution of cognition — A hypothesis." *Cognitive Science*, Vol. 27, pp. 135–155, 2003.

[4] G. Gallup, J.Anderson, and D. Shillito. "The Mirror Test." In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*, edited by M. Bekoff, C. Allen, and G. Burghardt. Cambridge, MA : MIT Press, 2002.

[5] D. Premack and G. Woodruff. "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, Vol. 4, pp. 515–526, 1978.

[6] C. Prince. "SoDiBot: Self-Other Discrimination Robot." http://www.d.umn.edu/˜cprince/PubRes/SoDiBot04/

[7] P. Rochat and T. Striano. "Who's in the mirror? Self-other discrimination in specular images by four- and nine-month-old infants." *Child Development*, Vol. 73, Number 1, pp. 35-46, 2002.

[8] P. Rochat. "Self-perception and action in infancy." *Experimental Brain Research*, Vol. 123, pp. 102–109, 1998.

[9] P. Rochat and R. Morgan. "Spatial determinants in the perception of self-produced leg movements by 3- to 5-month-old infants." *Developmental Psychology*, Vol. 31, pp. 626–636, 1995.

[10] M. Schmuckler. "Visual-proprioceptive intermodal perception in infancy." *Infant Behavior and Development*, Vol. 19, pp. 221–232, 1996.