# Learning About the Self and Others Through Contingency

**Kevin Gold** and **Brian Scassellati**
Department of Computer Science
Yale University
New Haven, CT, USA
kevin.gold@yale.edu and scaz@cs.yale.edu

### Abstract

Knowledge of the time delay between a robotic action and the reaction in the environment can be used to perform self-recognition and possibly identify other social agents in the environment as well. We describe a self-recognition system that uses timing information to identify self-generated motion and describe how this method might be extended to identify nonmoving parts of the robot. In addition, we show that our means of self-recognition can be applied to mirror self-recognition. Finally, we describe our plans to use contingency to identify other agents in the environment, using a method similar to that employed for self-recognition.

## Introduction

The ability to act upon the world provides a strong advantage in learning about it. Paul Fitzpatrick, for instance, has recently shown that a robot can learn a great deal of information about an object by reaching out and giving it a push (2003). In that work, the robot had to first find its own end-effector in the visual field, and it did this by moving its hand back and forth for a moment; the resulting motion identified which object in the scene was the robot.

In that study, this hand-waving behavior was a hard-coded action, performed before each trial to get a fix on the robot's hand; it was not a focus of the research itself. Nevertheless, temporal correlations such as this one may be of fundamental importance in learning about the self and others. While other recent robotics research on learning about the self has focused on the temporally invariant visual properties of the body (Yoshikawa et al. 2004b), a more general heuristic might be that the physical self is defined in two ways: that which can be immediately controlled, and that which can experience sensations. We shall focus on the former property here, though another group is making progress on the latter (Yoshikawa, Hosoda, and Asada 2004a).

When a robot acts and senses a response, the promptness of that response should indicate what kind of entity made it. If the response is nearly immediate, it is most likely sensing its own effectors, or (in the case of head movement) a perceptual change resulting from its action. If the response is delayed, or if it continues for a bit after the agent has stopped acting, the agent may be sensing a result in the environment of its action – for example, a ball the robot has dropped will continue to fall. If the response is delayed still further, the entity is likely to be a social agent. Humans can be shown to apply this heuristic subconsciously: when watching a movie of one circle rolling up to another, the delay between contact and the second circle rolling away dictates whether the second circle is seen as self-propelled (Michotte 1963). Thus, using the time delay between action and reaction is not a method specific to self-recognition, but a general means of structuring knowledge about a dynamic world.

The Yale Social Robotics Lab has begun work on a robotic time-delay-based system for recognizing and learning about the self and others. This work is being developed on Nico, an infant-like humanoid robot. Nico learns through experimentation to expect motion in its visual field within a certain time window after initiating an arm motor movement. Because we hope to understand how self-recognition could work in the absence of a priori knowledge of the body, the method does not make use of a kinematic model – though Nico may later learn such a model through experimentation. This absence of a kinematic model has the advantage that Nico can recognize self-generated motion in an unfamiliar context. For instance, when a mirror is placed in front of Nico, Nico can recognize the reflection of its self-generated motion in the mirror as easily as it can find its real end-effector.

Our research is novel in that it is able to recognize moving parts belonging to the robot in the visual field even in the presence of distractors. Similar systems have either assumed the absence of distractors (Chella, Frixione, and Gaglio 2003), made cause-and-effect predictions in a simpler sensory domain (Provost, Beeson, and Kuipers 2001), or assumed that the robot's form was static against a changing background (Yoshikawa et al. 2004b). Each of these systems is powerful in its own domain; however, none has solved the general problem of learning to recognize the robot's own moving parts in a real-world environment with distractors.

Below, we summarize our results so far and planned extensions in the domain of self-recognition, as well as our plans to employ a similar system for social agent recognition.

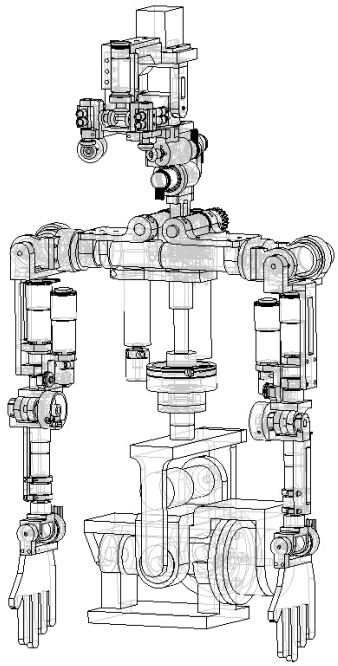


Figure 1: Line drawing of the robot's current mechanical design.

## 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 characteristics. 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.

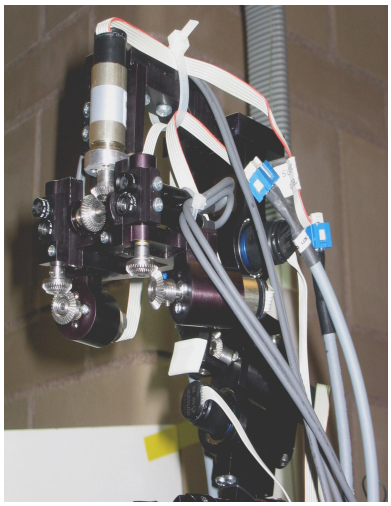


Figure 2: The robot's head-neck assembly, housing a four camera vision system (center, left) and providing a total of 7 degrees of freedom.

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 suffi-

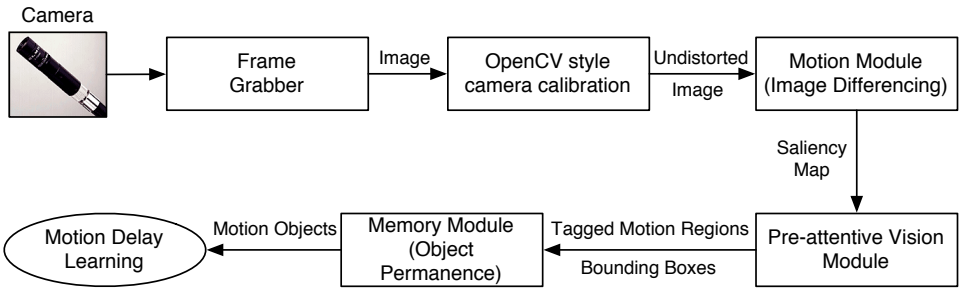



Figure 3: Visual and attentive processing preceding the delay learning stage. A separate processing flow is associated with each eye.

ciently similar, they are considered as corresponding to the same moving object and given the same object identifier.

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.

## Self-Recognition

In the visual domain, Nico learns to recognize self-generated motion by making random arm movements when left alone. For each arm movement, an event clock is initialized to 0 when the motor command is sent. When visual feedback is received in the form of a new motion bounding box, that time is recorded as $t_1$. (The times are recorded with QNX's realtime clock, which provides temporal resolution on the order of nanoseconds.) This behavior eventually produces a range of acceptable onset times of self-generated motion, $[t_{1min}, t_{1max}]$.

Though perfect perception would allow us to use this range directly, in practice it is possible to generate values of $t_1$ that are either too small, in the case of camera noise detected as motion, or too large, in the case of subtle wrist movements that do not generate sufficient motion to be detected at a reasonable time. For this reason, Nico keeps only the middle 95% of the $t_1$ values generated. This means that the aforementioned subtle movements may not be labeled accurately at test time – but this tradeoff is necessary to keep the time window meaningful with the introduction of distractors.

Fig. 4 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. Note, however, that the perceptual loop does not function as quickly as that of a human; the average $t_1$ hovers around 600 milliseconds. Nor is the detection of motion nearly as consistent, as the range of $t_1$ values produced is almost 400ms. These facts suggest the need for caution in applying our intuitions about human perception to the problem of robotic self-recognition.

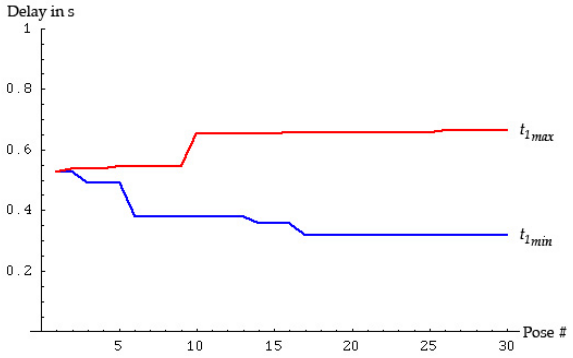
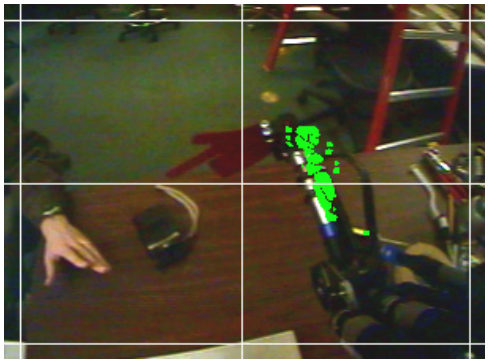



Figure 4: Evolution of the learned bounds on the self-recognition time window.

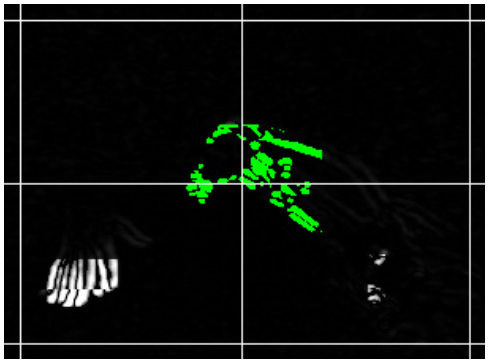
Once the $t_1$ window is established, it can be used to identify self-generated motion. Motion bounding boxes that first appear within the specified time window are labeled as "self", with the corresponding pixels indicated in bright green in the image output. An example of a successfully labeled image is shown in Fig. 5.

Currently, the system performs moderately well in the presence of distractors. To evaluate recognition accuracy in the presence of human-induced motion in the visual scene, we instructed an independent subject to immediately shake her hand in front of Nico in response to Nico's movement (Fig. 6). Out of 44 trials of this, some part of the subject's hand was mislabeled as self-motion only 15 times, yielding a 34% false-positive rate even under these strenuous circumstances. These results are probably not quite as good as previously reported for this method (Michel et al. 2004) on account of the use of an independent subject, who eliminated experimenter bias in the timing of the wave.

Much of the error and variability in the $t_1$ window can be attributed to the inherent noisiness of the motion data. Because optical flow methods are slow, we used simple thresholded image-differencing to obtain the locations of pixels where motion occurred. This means that motion was generally detected only at the edges of Nico's arm, where the differences between frames were greatest. If Nico's black arm passed in front of a dark shadow, some motion might be thresholded away entirely. Thus, the motion boxes are



(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).

Figure 6: 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.

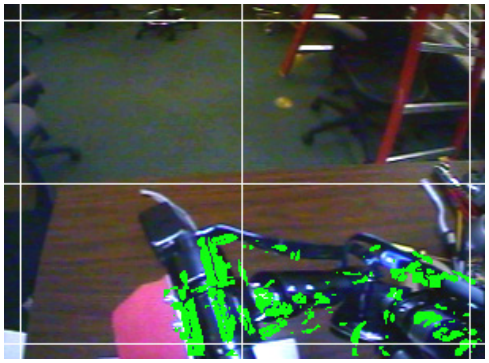


Figure 5: Output from the self-motion classifier, overlaid onto the visual input from one eye. All salient pixels from a moving object identified as 'self' are highlighted.

a somewhat unstable foundation for self-identification. An underlying object model with which the various motion patches could be associated would greatly increase the reliability of the method.

## Extensions to Self-Recognition

Now that we have a way of creating bounding boxes around self-generated motion, we can begin to extract properties about the self. The most obvious is color; by recording the pixel values within the general outlines provided by the motion module, we can generate a color histogram for the self. Region-growing should then provide a more stable referent for the arm than the motion bounding boxes, allowing us to refine the $t_1$ window further.

A stable referent for the arm will also allow us to learn its kinematics. Another member of our lab has implemented a system that learns a forward kinematic model, given a salient end-effector that the system can track (Sun and Scassellati 2004). Ideally, we would like that end-effector's appearance to be learned, rather than given a priori.

The condition that the arm begin its motion from within the field of view is somewhat restrictive, since Nico's peripheral vision is not as good as that of a human. This requirement could be relaxed if the robot learned to look first to its arm's natural rest pose, which it could do by finding a neck position for which the arm's time delay $t_1$ is minimized.

Introducing the neck motors brings up the question of how the robot would distinguish viewpoint changes caused by head turns from genuine self-motion. A fully satisfactory solution would require pulling the time delay method out of the low-level visual domain, and into an object model space that is left unchanged by viewpoint rotations. Changes in such a space that are contingent on the robot's motor commands could then be accurately labeled as "self." It is an interesting question as to how much a priori knowledge would actually be necessary to build such a space. Meanwhile, the question can be skirted by assuming a priori knowledge of which motors are viewpoint-changing; the human ocular system takes a similar approach in masking motion generated by saccades (Hubel 1995).

Finally, an ability to grow the self-label to nonmoving parts would allow Nico to recognize its whole image in the mirror. Currently, Nico would not be able to pass the "mirror test" performed on chimpanzees (Gallup, Anderson, and Shillito 2002) and infants (Rochat and Striano 2002) to ascertain self-awareness, because Nico only labels its currently moving parts as "self" (Fig. 7). Extending the self-label to the entirety of its body would allow Nico to associate self-touch with the self-image in the mirror, possibly using a method similar to that described in (Yoshikawa, Hosoda,

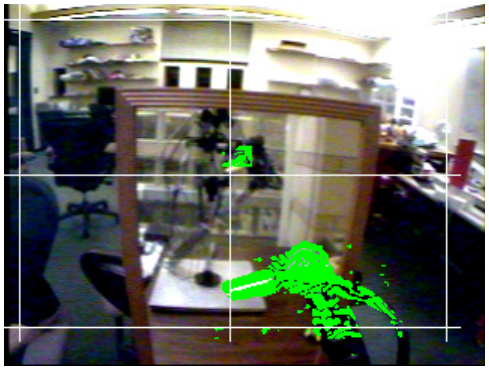


Figure 7: Nico recognizes self-motion in a mirror.

and Asada 2004) for non-reflected self images.

## Social Agent Recognition

Since gestures are not reliably produced in social interactions, it may prove difficult to use motion information to establish a time window for identifying social agents. Thus, vocalizations may be a more effective means of learning about other social agents.

Long before infants are able to make meaningful gestures, they are able to elicit a social signal by crying. We believe that in addition to signaling hunger or unhappiness, crying may be a means of sending a signal that elicits a social response, thus allowing the infant to learn about social beings. By remembering the visual properties of a caregiver summoned by vocalization, infants may in this way gain their first knowledge of the properties of other social beings. In the coming months, we will implement this behavior on Nico to determine what sorts of properties about the caregiver can be reliably learned through crying to get attention.

Moreover, we may be able to apply our method of learning a specific time window to the vocal domain as well. Javier Movellan has demonstrated a method (Movellan 2004) by which a robotic entity can ascertain through vocalizations whether it is interacting with a human contingently. Like our work, it categorizes input occuring within a first time interval as ”self” – in this case, the robot's own vocalization – and that occuring within a second window as indicating a contingent response. His method relies on a priori assumptions about the timing and duration of the relevant time windows and the probability distributions associated with them; we would like to determine with what degree of accuracy this kind of model can be learned from experience, and whether it can aid in identifying agents in the visual field through sound localization.

## Conclusions

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. In the coming months, we hope to determine how reliably other visual properties of the self can be extracted using this basic timing information.

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 perceptually grounded meanings for such difficult linguistic concepts as “I” and “you”. The grounding of abstract symbols and predicates in sensory data is known as the anchoring problem (Coradeschi and Saffiotti 2000). Most anchoring research has focused on anchoring objects external to the robot; and indeed, in many applications the robot has no need of an explicit concept of “self”. However, social interactions make heavy use of the concepts of “I” and “you”, making the ability to symbolically reason about these concepts desirable.

By associating a second characteristic time window with humans' reactions to its movements, Nico could also learn to 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).

## References

Chella, A.; Frixione, M.; and Gaglio, S. 2003. Anchoring symbols to conceptual spaces: the case of dynamic scenarios. *Robotics and Autonomous Systems* 43, p. 175-188.

Coradeschi, S.; and Saffioti, A. 2000. Anchoring symbols to sensory data: preliminary report. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2002)*, Menlo Park, Calif.: AAAI Press.

Fitzpatrick, P. 2003. From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot. PhD diss., Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Gallup, G.; Anderson, J.; and Shillito, D. 2002. The Mirror Test. In Bekoff, M.; Allen, C.; and Burghardt, G. *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition.* Cambridge, Mass.: MIT Press.

Hubel, D. 1995. *Eye, Brain, and Vision.* New York: Scientific American Library.

Michel, P., Gold, K., and Scassellati, B. Motion-Based Robotic Self-Recognition. In IEEE/RSJ International Conference on Intelligent Robotics and Systems, Sendai, Japan.

Michotte, A. 1963. *The perception of causality.* New York: Basic Books.

Movellan, J. 2005. Infomax Control as a Model of Real Time Behavior: Theory and Application to the Detection of Social Contingency, Technical Report, MPLab TR 2005-1, Machine Perception Laboratory, University of California, San Diego.

Provost, J.; Beeson, P.; and Kuipers, B. 2001. Toward learning the causal layer of the spatial semantic hierarchy using SOMs. In Proceedings of the AAAI-2001 Spring Symposium on Learning Grounded Representations, Palo Alto, CA.

Rochat, P.; and Striano, T. 2002. Who's in the mirror? Self-other discrimination in specular images by four- and nine-month-old infants. *Child Development* 73(1):35-46.

Sun, G.; and Scassellati, B. 2004. Reaching through Learned Forward Model. In Proceedings of IEEE-RAS/RSJ International Conference on Humanoid Robots, Los Angeles, CA.

Yoshikawa, Y.; Hosoda, K.; and Asada, M. 2004a. Cross-anchoring for binding tactile and visual sensations via unique association through self-perception. In Proceedings of the International Conference on Learning and Development, San Diego, CA.

Yoshikawa, Y.; Tsuji, Y.; Hosoda, K.; and Asada, M. 2004b. Is it my body? Body extraction from uninterpreted sensory data based on the invariance of multiple sensory attributes. In IEEE/RSJ International Conference on Intelligent Robotics and Systems, Sendai, Japan.