Using Probabilistic Reasoning Over Time to Self-Recognize

Kevin Gold and Brian Scassellati

Department of Computer Science, Yale University, 51 Prospect St., New Haven, CT, USA

Abstract

Using the probabilistic methods outlined in this paper, a robot can learn to recognize its own motor-controlled body parts, or their mirror reflections, without prior knowledge of their appearance. For each item in its visual field, the robot calculates the likelihoods of each of three dynamic Bayesian models, corresponding to the categories of "self," "animate other," or "inanimate." Each model fully incorporates the object's entire motion history and the robot's whole motor history in constant update time, via the forward algorithm. The parameters for each model are learned in an unsupervised fashion as the robot experiments with its arm over a period of four minutes. The robot demonstrated robust recognition of its mirror image, while classifying the nearby experimenter as "animate other," across 20 experiments. Adversarial experiments in which a subject mirrored the robot's motion showed that as long as the robot had seen the subject move for as little as 5 seconds before mirroring, the evidence was "remembered" across a full minute of mimicry.

Key words: self-recognition, robot, mirror test, dynamic Bayesian model, animacy, contingency

1 Introduction

This paper presents a simple algorithm by which a robot can learn over time whether an item in its visual field is controllable by its motors, and thus a part of itself. Because the algorithm does not rely on a model of appearance or even kinematics, it would apply equally well if the robot were damaged, moved into different lighting conditions, or otherwise changed its appearance.

Email address: kevin.gold@yale.edu, scaz@cs.yale.edu (Kevin Gold and Brian Scassellati).

Perhaps more compelling is the fact that the method applies equally well to the robot's own unreflected parts and its reflection in the mirror.

Much has been made of mirror self-recognition in animals and humans, and some psychologists are quite willing to interpret mirror self-recognition as evidence for a sense of self [13,5]. The use of a mirror to assay intelligence is partly popular because the dividing line seems to so clearly segregate the intelligent species from the not-so-intelligent: among animal species, only apes [6,7], dolphins [18], and elephants [16] are known to learn in an unsupervised manner that their mirror reflections are themselves, while monkeys treat their reflections as conspecifics [11]. Human children typically begin to pass the "mirror test" around the age of two [1], which is about the same time they begin to use personal pronouns [3]. However, just as Deep Blue's chess-playing does not necessarily imply well-rounded intelligence, mirror self-recognition in a robot cannot necessarily be interpreted as evidence for more general cognitive abilities. Even Gray Walter's mechanical tortoises of the 1950s displayed different behavior in front of a mirror than not, but this was a simple consequence of the robot's alternating attraction and repulsion from its own light source [21]. Furthermore, the evidence assigning any importance to mirror self-recognition even among animals and humans is at best suggestive. We shall therefore avoid the hyperbolic claims of, e.g., [20] that our robot is "conscious." We claim only that it can learn to reliably distinguish its own moving parts from those of others; any language used in this paper that suggests any kind of agency on the part of the robot should be taken to be only offered as analogy.

As an implementation of robotic self-recognition based on motion or timing, the method has several advantages over previous work. The most crucial is that unlike [14] and [10], the present method takes into account the whole observation history of an object, rather than only reacting to its current motion or state. This makes the algorithm more resistant to noise, more consistent over time, and able to remember that objects temporarily moving simultaneously with the robot are not actually itself. The current method is also more transparent than previous methods such as [20], which used a recurrent neural network to produce different behavior in front of a mirror than not. The present method produces explicit probabilities for each classification, using probabilities and calculations that themselves have intuitive semantics, and thus simplifies the task of interpreting what the robot is actually calculating. Finally, other researchers have described methods that simply produce different behavior in front of a mirror, rather than any representation that is accessible for further probabilistic reasoning [20,10,21]. Because our method produces probabilities with clearly defined semantics, the results can be more easily integrated with the robot's other mechanisms for probabilistic reasoning.

The algorithm compares the likelihoods of three dynamic Bayesian models at every moment in time. One model corresponds to the hypothesis that the robot's own motors generated an object's motion; the second model corresponds to the hypothesis that something else generated that motion; and a third model detects irregular motion, such as that caused by noise or dropped inanimate objects. Given the history of visual evidence for whether an object has moved from frame to frame, and the kinesthetic evidence for whether the robot's own motors were moving at a particular frame, it is possible to calculate a probability for each of these models for a particular object, and update these probabilities in constant time. If the robot can consistently control something's motion, then that thing is considered to belong to the robot's own body.

Other methods of robotic self-recognition have not relied on motion, and thus have come with their own advantages and drawbacks. A robot can, for instance, touch itself and compare its visual feedback to its haptic feedback, thereby creating a visual-somatosensory map [23]. This method obviously requires the recognized areas to possess touch sensors and be reachable, but as an advantage over the present method, the recognized areas would not need to be motor-controlled. Another method is to statistically extract parts of the visual scene that remain invariant in different environments [22]. This does not work well for either moving parts or mirror images, but could detect parts of the robot that move with the camera. A third method is to find salient patches of the visual scene, cluster them over time by their color histograms, and determine which clusters' positions have the highest mutual information with the robot's kinematic model [12]. This method creates and relies on expectations for the appearance and position of the robot's parts, which may work less well for identifying parts under transformations such as mirror reflection or changed lighting conditions, but could be useful in bootstrapping a forward model for reaching.

Section 2 describes the underlying mathematical model that produces the classification probabilities. Section 3 describes how the model was implemented on the humanoid upper-torso robot, Nico (Figure 1). Section 4 describes the results of experiments in which the robot learned the parameters of its self model by watching its own unreflected arm for four minutes, and then classified its mirror image and the experimenter. Section 5 describes experiments in which a human adversary mirrors the robot's motion. We conclude with some speculations about the significance of the "mirror test" as a test of intelligence, some hypotheses about how mirror self-recognition might function in the wild, and some critiques and further extensions of our method. (Sections 2–4 appeared in abbreviated form in a proceedings paper for the Cognitive Science Society [9], while the experiments in Section 5, this introduction, and the conclusions are new to this paper.)

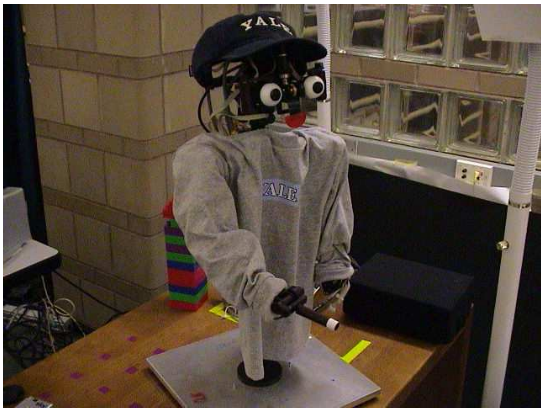

Fig. 1. Nico is an upper-torso humanoid robot with the arm and head kinematics of a one-year-old.

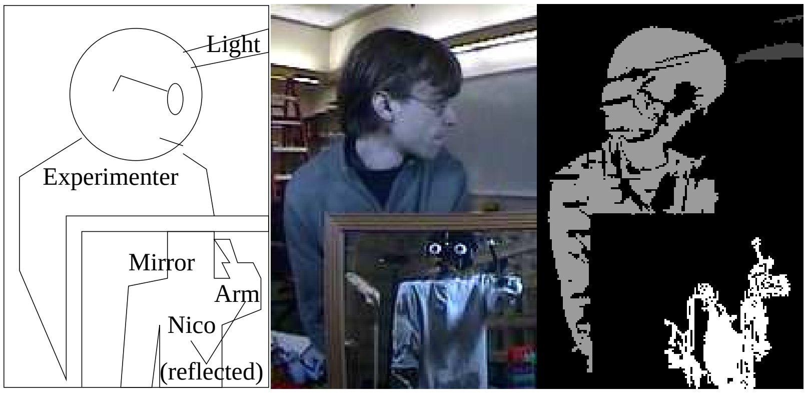

Fig. 2. Output from the self/other algorithm (bottom) on the video streaming from the robot's camera (upper right). The robot classifies its mirror image as "self" (white), the experimenter as "animate other" (light gray), and the image noise from the overhead light as "inanimate" (dark gray). Objects in this implementation were found through background subtraction, thus excluding the nonmoving parts of the robot.

2 Mathematical Background and Models

Our method compares three models for every object in the robot's visual field to determine whether it is the robot itself, someone else, or neither. The use of Bayesian networks allows the robot to calculate at each time t the likelihoods λ_t^{ν}, λ_t^{σ}, and λ_t^{ω}, corresponding to the likelihoods of the evidence given the inanimate model, the self model, and the "animate other" model, respectively. Normalizing these likelihoods then gives the probability that each model is correct, given the evidence. We shall first discuss how the models calculate their probabilities under fixed parameters, then explain how the parameters themselves are adjusted in real-time.

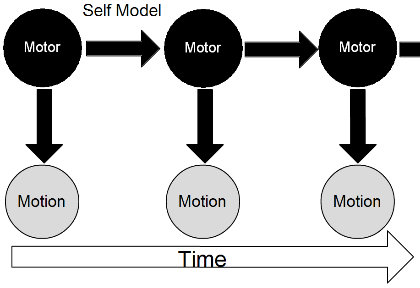

Fig. 3. The robot's self model, in graphical model notation. Arrows represent conditional dependence relations, and circles represent events. The likelihood of the observed motion evidence (gray) is calculated conditioned on the robot's knowledge of its motor states (black).

The "inanimate" model is the simplest, as we assume inanimate objects only appear to have motion due to sensor noise or when they are dropped. If we characterize the occurrence of either of these events as the event r, then this model is characterized by a single parameter: the probability $P(r)$ that random motion is detected at an arbitrary time t. Observations of this kind of motion over time are assumed to be independent, such that the overall likelihood λ_t^{ν} can be calculated by simply multiplying the likelihoods at each time step of the observed motion.

The robot's second model for an object is the "self" model, in which the motor actions of the robot generate the object's observed motion. The model is characterized by two probabilities: the conditional probability $P(m|\phi)$ of observing motion given that the robot's motors are moving, and the conditional probability $P(m|\neg\phi)$ of observing motion given that the robot's motors are not moving. (Henceforth, m and $\neg m$ shall be the observations of motion or not for motion event M, and ϕ and $\neg\phi$ shall serve similarly for motor event Φ. Note that these probabilities need not sum to 1.)

Figure 3 shows the graphical model corresponding to the robot's "self" model. Each circle corresponds to an observation of either the robot's own motor action (top circles) or the observed motion of the object in question (bottom circles), with time t increasing from left to right. The circles are all shaded to indicate that these event outcomes are all known to the robot. The arrows depict conditional dependence; informally, this corresponds to a notion of causality. Thus, the robot's motor action at time t causes the perception of motion at time t.

To determine the likelihood of this model for a given object, the robot must calculate the probability that its sequence of motor actions would generate the observed motion for the object. The relevant calculation at each time step is the probability $P(M_t|\Phi_t)$ of motor event Φ_t generating motion observation M_t. These probabilities calculated at each time step can then be simply multiplied together to get the overall likelihood of the evidence, because the motion observations are conditionally independent given the robot's motor actions. [1]

The likelihood λ_t^σ that the motion evidence up to time t was generated by the robot's own motors is then:

$$\lambda_t^\sigma = \prod_t P(M_t|\Phi_t) \tag{1}$$

where, in our simple Boolean implementation,

$$P(M_t|\Phi_t) = \begin{cases} P(m_t|\phi_t) & \text{if } m_t \text{ and } \phi_t \\ 1 - P(m_t|\phi_t) & \text{if } \neg m_t \text{ and } \phi_t \\ P(m_t|\neg\phi_t) & \text{if } m_t \text{ and } \neg\phi_t \\ 1 - P(m_t|\neg\phi_t) & \text{if } \neg m_t \text{ and } \neg\phi_t \end{cases} \tag{2}$$

Under this model, updating the likelihood at time $t+1$ is simply a matter of multiplying by the correct value of $P(M_{t+1}|\Phi_{t+1})$:

$$\lambda_{t+1}^\sigma = P(M_{t+1}|\Phi_{t+1})\lambda_t^\sigma \tag{3}$$

Thus, the robot need not maintain the whole graphical model in memory. Only the likelihood calculated at the previous time step $t-1$ needs to be retained to calculate the likelihood at the current time step t.

Note that equation 1 and the graphical model presented in Figure 3 are much more general than the simple Boolean model implementing them that is described by equation 2. For more advanced models, M_t could be a complete reading of joint angles, Φ_t could describe a trajectory through space, and $P(M_t|\Phi_t)$ could be an arbitrary distribution on motion trajectories given the motor readings. The current implementation, however, chooses simplicity over expressive power.

The third and final model is that of another person (or other animate agent) in the visual field. This model is identical to the self model, but now the

[1] Though the graphical depiction of the self model includes the conditional dependence relations of the robot's activity from one time step to the next, these transitions do not actually matter for the calculation of the likelihood of the motion evidence, which is *conditioned on* the motor evidence. When conditioning on event nodes, conditional dependence between them becomes irrelevant. We include the motor dependence arrows here to better illustrate the point that the "animate other" model is exactly the "self" model, with only a change in what evidence is assumed to be available; but we could as easily omit them, as we do in [9].

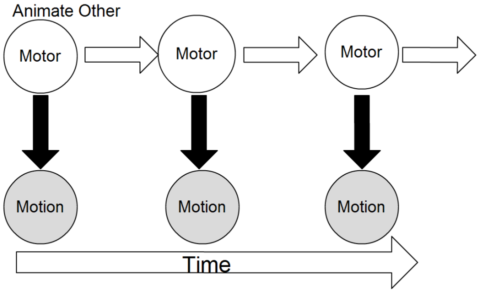

Fig. 4. The model for an "animate other," in graphical model notation. The model the robot uses is a copy of its self model, but with the motor information unobserved and inferred (unshaded circles).

motor states are hidden to the robot, leaving it to infer the other person's motor states. Removing the motor information turns the model into a Hidden Markov Model (HMM; Figure 4). To assess the likelihood that motion was generated by another animate agent, the robot must now infer the underlying motor states that generated the motion. Performing this calculation requires two transition probabilities, $P(\phi_{t+1}|\neg\phi_t)$ and $P(\neg\phi_{t+1}|\phi_t)$, corresponding to the probabilities that the person begins motor activity from rest or ceases its motor activity, respectively.

To find the likelihood that an object is an animate other, the robot must keep track of the probabilities of each motor state at each time step. This can be updated online in a constant amount of time at each time step using the forward algorithm [17]. The forward algorithm can be described by the following equation for calculating the new likelihood λ^{ω}_{t+1} given an observation of motion M_{t+1} and motor state probabilities $\overrightarrow{P(\Phi)}$:

$$\lambda^{\omega}_{t+1} = \sum_{\Phi_{t+1}} P(M_{t+1}|\Phi_{t+1}) \sum_{\Phi_t} P(\Phi_{t+1}|\Phi_t) P(\Phi_t) \tag{4}$$

Notice that this equation necessarily entails calculating probabilities for the entity's hidden state as a subroutine. In our simple implementation, this only decides whether the agent is engaging in motor activity or not. But, as with the self model, equation 4 and figure 4 are more general than the simple boolean model we are using here. A more complex model relating the entity's motor states to its motion would allow action recognition as a pleasant side effect of the likelihood calculation.

The normalized likelihood for each category can be considered to be the probability that an object belongs to that category; thus $P_t(\text{self}) = \lambda^{\sigma}_t/(\lambda^{\sigma}_t + \lambda^{\omega}_t + \lambda^{\nu}_t)$, and so on. These normalized likelihoods can be propagated in the likelihood calculations instead of the original likelihoods without changing the ratios between them, and this avoids underflow. [2]

The forward algorithm requires prior probabilities on the possible motor states to propagate forward. Since the robot has no particular information about the state of the "other" at time 0, it arbitrarily sets these to 0.5. However, both the "self" and "other" models are more complex and more rare than the assumption of noise, so the final likelihoods are weighted by the priors of $P(\text{inanimate}) = 0.8$, $P(\text{self}) = 0.1$, $P(\text{other}) = 0.1$. As constants, these priors do not matter much in the long run, but they can reduce false classifications when first encountering an object.

The other parameters to the model, the conditional probabilities, do matter in the long term; luckily, they can be learned online rather than set arbitrarily. For the "inanimate" and "self" models, the robot does this by counting its observations of each event type ($\{m, \neg m\} \times \{\phi, \neg\phi\}$) and weighting each observation by the probability that the object truly belongs to the model for which the parameters are being calculated. Thus:

$$P(r) = \frac{\sum_{it} P_{it}(\text{inanimate}) M_{it}}{\sum_{it} P_{it}(\text{inanimate})} \tag{5}$$

$$P(m|\phi) = \frac{\sum_{it} P_{it}(\text{self}) M_{it} \Phi_t}{\sum_{it} P_{it}(\text{self}) \Phi_t} \tag{6}$$

$$P(m|\neg\phi) = \frac{\sum_{it} P_{it}(\text{self}) M_{it}(1 - \Phi_t)}{\sum_{it} P_{it}(\text{self})(1 - \Phi_t)} \tag{7}$$

where $P_{it}(\text{inanimate})$ is the robot's best estimate at time t of the probability that object i is inanimate, and M_{it} is 0 or 1 depending on whether object i is moving. This strategy is a kind of expectation maximization, because it alternates between fitting a model to data and classifying the data with a model. Normally, expectation maximization requires iterating over all previous observations in updating the model, but this model is simple enough that the robot can update it in real time without going back to revise its previous probability estimates, which will asymptotically contribute little to the overall likelihoods.

Since this method is iterative, the robot must begin with some estimate of each of the probabilities. The robot begins with a guess for each parameter as well as a small number of "virtual" observations to support that guess. These guesses function as priors on the parameters, as opposed to the priors on classifications described earlier. The choice of priors here does not matter much, since the

[2] If this is done, the state probabilities of the "animate other" case must be scaled to sum to $P(\text{animate})$ before propagation; as shown in Equation 4, λ_t^ω does not itself figure into the calculation of λ_{t+1}^ω except by way of these state probabilities.

system can adapt to even bad priors (see "Experiments," below). Any prior will smooth the model's development of the correct parameters, by reducing its reliance on its first few observations.

Using the expectation maximization strategy on the "animate other" model would not work quite as well, because it contains unobserved states. Technically, to perform expectation maximization on a Hidden Markov Model requires the forward-backward algorithm [2] to obtain *a posteriori* estimates of the hidden states, which would require iterating over all the data repeatedly as the robot gained more data. However, we can finesse this problem, reduce the size of the hypothesis space, and prove an interesting point about self-models all at the same time if the robot uses its own self model to generate the "animate other" model. The probabilities $P(m|\phi)$ and $P(m|\neg\phi)$ are set to exactly the same values as the self model; this is equivalent to the assumption that the robot has about the same chance of perceiving motion if either itself or someone else is actually moving. The transitional probabilities $P(\phi_{t+1}|\neg\phi_t)$ and $P(\neg\phi_{t+1}|\phi_t)$ are based on the robot's own motor activity by counting its own action transitions of each type. Though the human's motions are likely to be quite different from those of the robot in their particulars, the general fact that "objects in motion tend to stay in motion" is true of both, and the real discrimination between the two hinges on the contingency with the robot's own motor actions, and not the particular transition probabilities of the "animate other" model.

3 Robotic Implementation

The following details are not necessary to the generic self-recognition algorithm, but describe the particular implementation on Nico, the robot that was used for our experiments.

Nico is a small upper-torso humanoid robot with the proportions and kinematics of a one-year-old child (Figure 1). Nico possesses an arm with 6 degrees of freedom, corresponding to the degrees of freedom of a human arm up to the wrist. These experiments used only motors for the elbow and two degrees of freedom at the shoulder; this excluded the two degrees of freedom at the wrist and the shoulder motor that would lift the arm out of view to the right. The wrist motors were excluded because their effects tend to be difficult for background subtraction to see. The single shoulder motor was excluded to expedite the learning, since the arm would otherwise spend most of its time outside the camera's field of view.

The arm made sweeping gestures roughly 1 second in length to a randomly chosen position roughly every five seconds. The arm swung in the forward direction between 45 and 70 degrees from vertical, rotated inward 0–20 degrees from the robot's facing direction, and bent at the elbow 0–80 degrees. New positions were chosen at random from among the extremal positions. Each movement took roughly 1 second to perform. Feedback from the motors in the form of optical encoder readings indicated to the robot whether each motor had stopped moving.

For vision, Nico used 320×240 images pulled from the wide-angle CCD camera in Nico's right eye at roughly 30 frames per second. Nico's head remained in a fixed position of 14 degrees below horizontal and 25 degrees to the right as it watched its right arm move during training. During testing, Nico looked seven degrees below vertical and 30 degrees to the left, so that the moving right arm would not get in the way of its view of the mirror. After Nico's head had moved to the correct position, the background subtraction filter used the subsequent 120 frames (4 seconds) to create a statistical model of the scene in which the mean and variance of each pixel's brightness was calculated. In processing the subsequent frames, pixels that did not vary from their mean by more than three standard deviations were discarded as background, leaving only the blobs of objects that had moved to their locations more recently and scattered pixel noise. Connected regions that did not exceed 100 pixels (roughly 0.1% of the image) were discarded as noise.

Objects were tracked over time by matching each region R_i in frame F with the region in frame $F-1$ that shared the largest number of pixels with R_i; regions of overlap were computed using four-connected depth-first search. If more than one connected region in the same frame attempted to claim the same object identity, as frequently happened when joined regions separated, a new identity was generated for the smaller region. An object with area A was judged to be moving if $4\sqrt{A}$ of its pixels had changed their region label from one frame to the next. This formula was chosen to be roughly proportional to the length of the object's perimeter, while taking into account that background subtraction tended to produce "fuzzy" borders that are constantly changing.

The final output of vision processing was an image of labeled regions that could be tracked over time and judged at each time step to be moving or not moving. This output was made available at a rate of roughly 9 frames per second. For each segmented region i for each frame t, the probabilities $P_{it}(\text{self})$, $P_{it}(\text{animate})$, and $P_{it}(\text{inanimate})$ were calculated and updated in real time using the algorithms described earlier. Figure 2 shows output typical of the self-other algorithm after learning, with image regions grayscale-coded by maximum likelihood classification. Note that because the background subtraction algorithm blinds the robot to objects that have not moved since the start of the experiment, the robot cannot actually classify its body, but only its arm. [3]

[3] In practice, a moving arm also creates lighting changes on the torso that can be

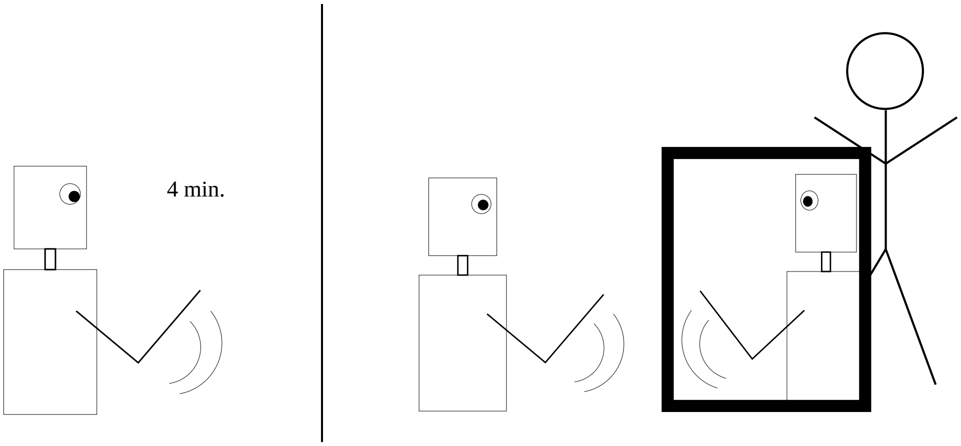

Fig. 5. The basic methodology for the recognition experiments. The robot experimented with its arm for four minutes (left), then had to judge whether its mirror image and the experimenter were itself (right). (Compare to Figure 2, showing the robot's labeled camera output under such test conditions.)

A segmentation algorithm based on depth would join the arm to the full body and classify the whole assembly based on the movement of the arm, but this was not implemented.

4 Recognition Experiments

4.1 Methodology

Figure 5 illustrates the learning and test phases of the recognition experiments. The robot was given 4 minutes to observe the movements of its own arm, starting with $P(r)$, $P(m|\phi)$, and $P(m|\neg\phi)$ all set to the implausible value of 0.5. These starting values were given the weight of 30 virtual observations, or roughly 3 seconds of data. The robot made its observations in the absence of a mirror and without any explicit feedback. Distractors from the arm included both inanimate objects (light fixtures that background subtraction had failed to remove) and animate others (students passing in the hall adjacent to the lab). To automatically collect data on the robot's hypotheses about its arm without hand-labeling each frame, the probabilities generated for the largest object within its field of view were recorded as data; this object was the arm in most instances.

The robot's parameters were then frozen at the four minute mark for testing, to ensure the robot's performance was based solely on its observations of its own unreflected arm. Using the parameters it learned during the previous

picked up by background subtraction.

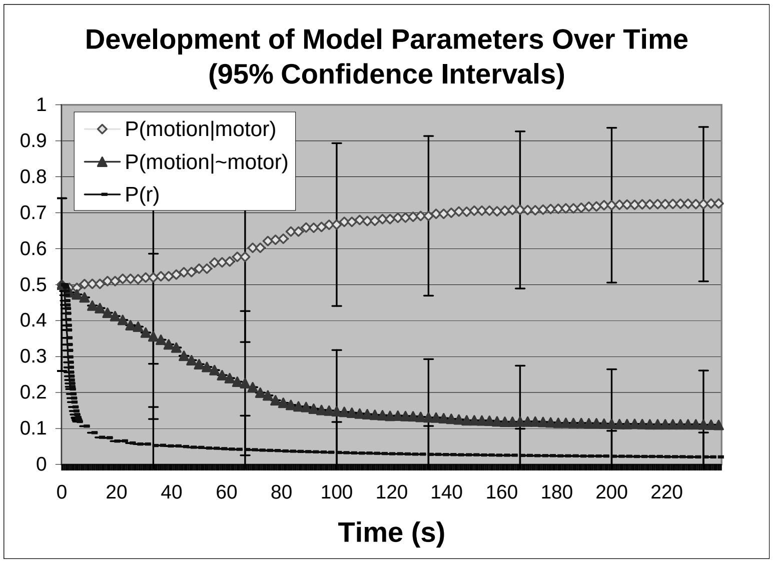

Fig. 6. Average model parameters over four minutes of unsupervised learning on the robot's visual and motor feedback, with 95% confidence intervals.

four minutes, the robot then was presented with a mirror, and the robot continued its random movements in front of the mirror. The robot's hypotheses for the largest object within the area covered by the mirror were recorded automatically, to avoid the chore of hand-labeling frames.

Using the same parameters, the robot then judged one of the authors (K.G.) for two minutes. Again, the robot's hypotheses for the largest object located within the area of interest were recorded automatically. The author's actions varied from near inactivity (sitting and checking his watch) to infrequent motion (drinking bottled water) to constant motion (juggling), while the robot continued to make periodic movements every 5 seconds. The visual feedback indicating the robot's current classifications was made unavailable to avoid biasing the experimenter's actions.

These experiments were repeated 20 times, to verify the robustness of both the learning mechanism and the classification schemes that it generated. Each learning trial reset all model probabilities to 0.5, and each test trial used the parameters generated in the corresponding learning trial.

4.2 Results

Within four minutes of learning, the robot's model probabilities consistently changed from 0.5 to roughly $P(r) = 0.020$, $P(m|\phi) = 0.73$, and $P(m|\neg\phi) = 0.11$. Figure 6 shows the mean values over time for these parameters, along with 95% confidence intervals calculated using Student's t for 19 degrees of freedom. The robustness of the model even under these starting conditions

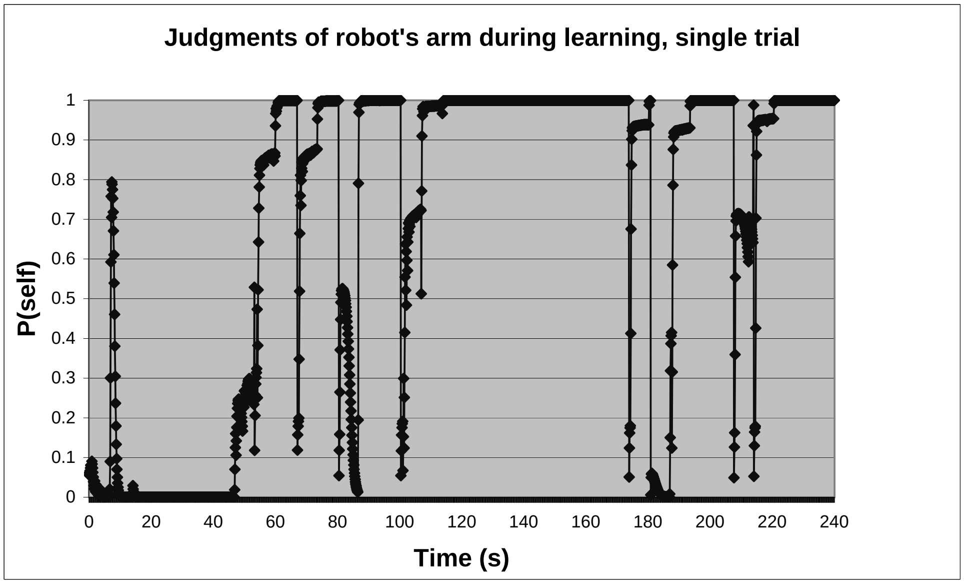

Fig. 7. Robot's judgments of its own arm during learning. A strong prior against "self" keeps this probability small at first. Later, the arm passes out of the robot's visual field several times, only to be quickly reclassified as "self" on its return.

was somewhat surprising; one would expect that the models would require some initial notion that motion was more likely with motor activity, but this fact was entirely learned.

This learning also occurred in the presence of many tracking failures, caused by the failure of background subtraction to correctly segment the image or the arm passing out of the field of view. Over each trial, the arm passed out of the field of view or was incorrectly segmented many times, only to have the "new" object quickly reclassified correctly from scratch (Figure 7).

When confronted with a mirror after the learning phase, the robot consistently judged the mirror image to be "self" after a single complete motion of its arm, and remained confident in that estimate over time. Figure 8 shows the robot's estimates of its mirror image over just the first 30 seconds, so as to better highlight the rapid change in its hypothesis during its first movement. The periodic dips in its confidence were not significant, but were probably caused by the slight lag between the robot sensing its motor feedback and seeing the visual feedback, as its motors needed time to accelerate to a detectable speed.

The robot's judgments of the experimenter as "animate other" were similarly quick, and its confidence remained high throughout every test trial. Figure 9 again shows only the first 30 seconds, to better highlight the changes in the first two seconds. The data for the next minute and a half was quite similar.

In short, the robot correctly classified both its mirror image and the experimenter quickly, persistently, and in all trials, using only the parameters it

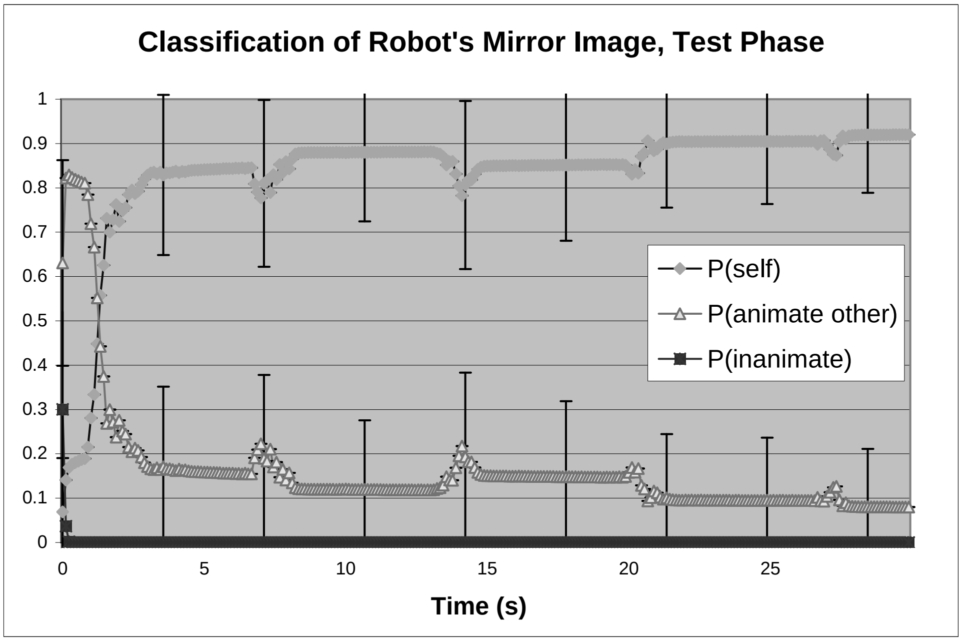

Fig. 8. Robot's judgements of its mirror image after 4 minutes of observing its own unreflected movements, with 95% confidence intervals.

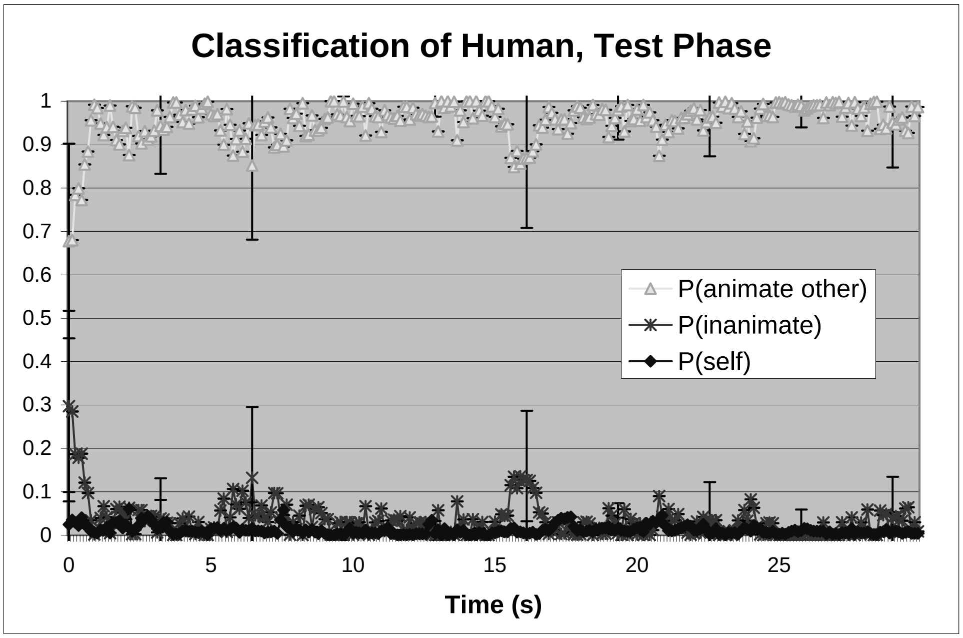

Fig. 9. Robot's judgments of the experimenter after 4 minutes of observing its own unreflected movements, with 95% confidence intervals.

learned from watching its unreflected arm for four minutes.

5 Adversarial Experiments

The recognition experiments described above were designed to test whether, under fairly natural conditions, the robot could learn appropriate parameters for its self model, and use these to correctly identify its mirror image as "self" while discounting the possibility that another human was itself. But because the system relies so much on the simultaneity of motion, it was an open question as to whether a person acting in tandem with the robot would be falsely identified as "self." Would the robot incorrectly assign a high probability of "self" to a human mirroring it?

There were two potential reasons to think the robot would not be deceived. First, it was possible that the temporal acuity of the robot was high enough that a human would not be able to respond quickly enough to match the robot's actions; the slight discrepancies, on the order of hundreds of milliseconds, would be enough to reveal that the human was actually someone else.

Second, because the forward algorithm propagates all information forward for an object from the time it is first seen, it was possible that the robot would be able to distinguish a mirroring human from itself because it had already seen the person move into place, in the absence of a corresponding motor signal, and that this information would propagate forward into the present, such that the robot would implicitly "remember" that earlier motion. (This is arguably a more natural scenario than mirroring the robot from the time it "wakes up.")

The next two experiments tested whether either of these reasons would be sufficient for the robot to distinguish a mirroring human from its mirror image on the basis of motion evidence alone.

5.1 Methodology

In the first adversarial experiment, a human subject was told to mirror the robot's motions as closely as possible, moving exactly and only when the robot moved (Figure 10). (The subject was unaware of the purpose of the experiment.) The subject stood roughly 1.8 meters away from the robot, in front of a plain cloth screen (for ease of segmentation). The subject stood in front of the screen during calibration for background subtraction as well, making her essentially invisible to the robot except for her moving hand.

The robot proceeded to make random arm movements every 5 seconds for a minute, using the parameters it had learned in the final run of the previous experiments to make online estimates of $P(\text{self})$ for each item in the visual field

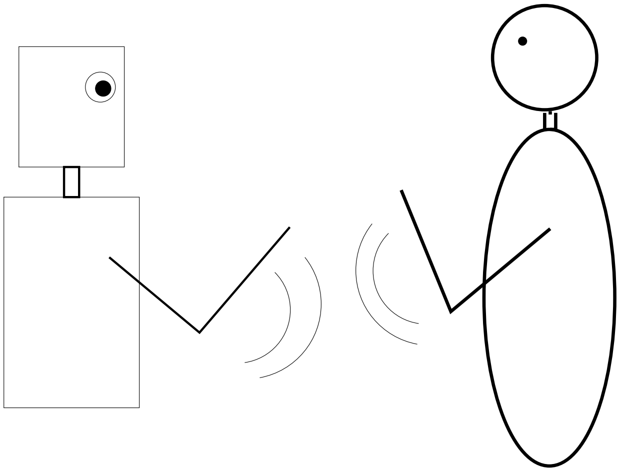

Fig. 10. The adversarial experiment, in which a human adversary mimicked the robot. In one condition, the robot was allowed to view the subject move into view; in another, the human imitated the robot from the time of its activation.

at a rate of roughly 8.3 Hz. Probabilities were recorded automatically for the largest segmented object within the region of the visual field delimited by the cloth screen, this being the human in almost all frames. The experiment was performed 10 times to determine 95% confidence intervals for the probabilities at each time.

The second adversarial experiment was identical to the first, except that the experiment now began with the subject off-camera. Each trial began with the subject spending the first 5 seconds or less moving into position in front of the robot, and the remaining 55 or more seconds mirroring the robot. This version of the experiment was repeated 5 times.

5.2 Results

During 9 out of the 10 runs in which the human mirrored the robot from the time the robot was turned on, the robot assigned a probability of “self” greater than 0.5 to the human at some point during the minute of imitation, suggesting that reaction time alone is not enough to dissuade the robot from being deceived by human mirroring. However, Figure 11 shows that the robot remained fairly uncertain about its judgments, and tended not to assign high probabilities of “self” to the human. (Compare this figure to the robot’s judgments about its own mirror image, Figure 8.) In fact, for fully half of the minute, the upper bound of the 95% confidence interval for the human’s $P(\text{self})$ value fell below 0.5.

When the robot was given the chance to see the human approach before mirroring, only in 2 of the 5 trials did the $P(\text{self})$ ascribed to the human exceed

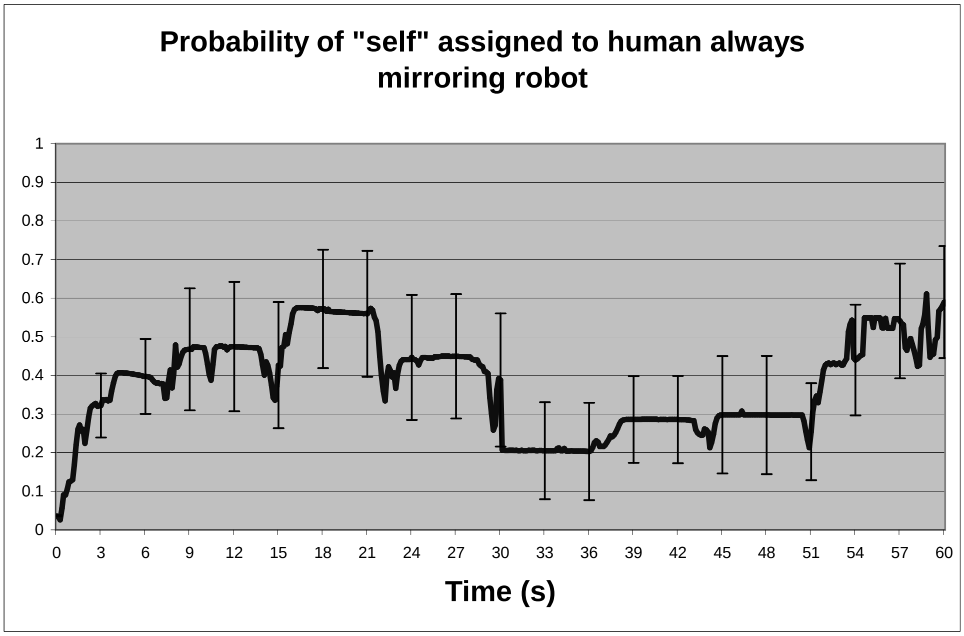

Fig. 11. Average $P(\text{self})$ attributed to a human mirroring the robot over the course of a minute, with 95% confidence intervals. The robot remains mostly uncertain of whether the mirroring human is itself, because it has never seen the human move except in time with its own motor activity.

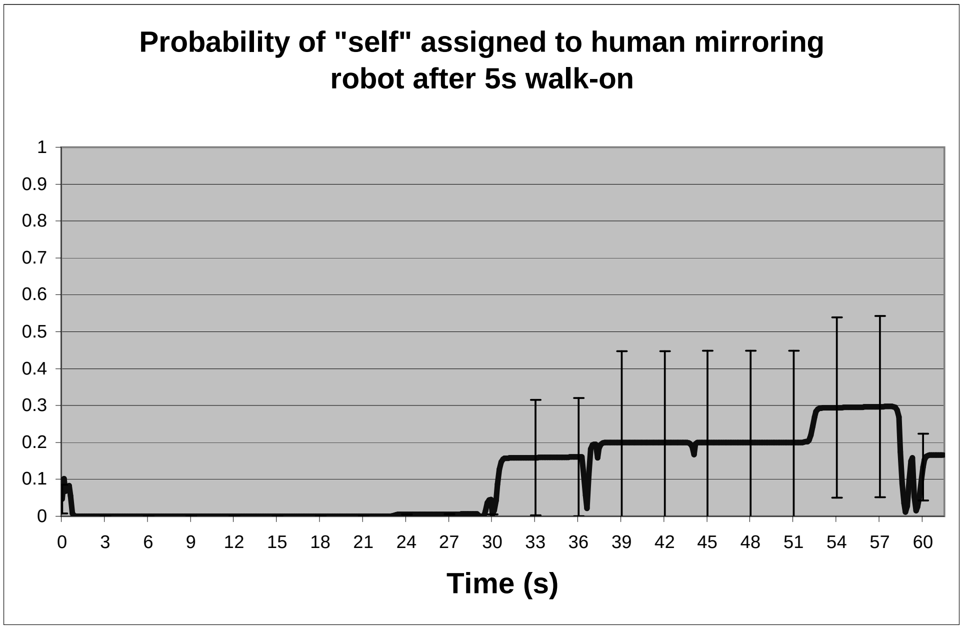

Fig. 12. Average $P(\text{self})$ attributed to a human mirroring the robot after a 5 second period in which the human walks into view. The evidence against "self" seen in the first 5 seconds persists throughout the mirroring, as it is implicitly propagated by the forward algorithm.

0.5. Figure 12 illustrates the average $P(\text{self})$ value ascribed to the human at each time step over the course of the minute-long experiment. Clearly, the evidence of the first 5 seconds tended to persist in affecting the robot's judgments. Only at the 30 second mark do the 95% confidence intervals for $P(\text{self})$ even become visible, because the deviations of the probability from 0 during the first half minute were quite small; during all 5 trials, $P(\text{self})$ remained below 0.1 during this time. In fact, for the 3 trials in which the robot never ascribed a high $P(\text{self})$ to the human, the "animate other" probability remained nearly 1 for the entire experiment; the climb at the end of the experiment in average values is due entirely to the two runs in which the human subject managed to successfully fool the robot. (This suggests that the confidence intervals in this experiment should be interpreted with caution, since the samples of the mean are probably not normally distributed.) In short, the motion evidence from the human's approach continued to persuade the robot that the human was not "self" long after the human had begun to mirror the robot.

6 Conclusions

The relatively simple Bayesian models we have suggested here are apparently sufficient to achieve self-recognition that is flexible enough to identify the robot's moving body parts in the mirror, yet robust enough to avoid misclassifying a human that begins to mirror the robot. We reiterate that the robot contains no programming specific to the case of the mirror; the mirror self-identification falls out naturally from the probabilistic reasoning about the cause of the motion in the mirror. Though the robot's classification is more ambivalent for any entity that perfectly mimics it from the time it is turned on, in general this is perfectly reasonable behavior: anything that the robot can control so effectively, probably may as well be classified as "self" until some evidence is acquired to the contrary.

Again, we reiterate that this result is an engineering solution for a robot, and not a model of human or animal mirror self-recognition. Yet, it would seem short-sighted to claim that absolutely *no* insight into mirror self-recognition in living species can be drawn from this experiment. One notices in nature that very different species that attempt to solve the same engineering problem can converge on similar mechanisms to solve it, through evolution. Though *mirror* self-recognition would never have evolved in living species, a general ability to learn to recognize what one has direct control over might have; and in attempting to replicate this ability in a robot, it is possible that we have hit upon some of the same solutions nature has. Therefore, we here very tentatively put forward some suggestions about what we might believe about human and animal mirror self-recognition, *if* it uses a mechanism that is at all similar to our solution.

The more we learn about intelligence, both natural and artificial, the more it becomes apparent that intelligence that functions in the real world is (and should be) fundamentally modular. A subsystem that performs self-recognition, to the exclusion of all else, is neither "cheating" at the mirror test nor hovering at the threshold of robot sentience. The idea that a single ability, be it mirror self-recognition, chess-playing, or performance on IQ-test analogies, can be the *ne plus ultra* test of intelligence – or "self-awareness," or any other vague but compelling capacity – does not even match what is known about human intelligence [8] or the modularity of the human brain [15], much less the history of AI. As it turns out, self-recognition is simply another domain of intelligence that is best solved by rather specific means, which should probably surprise no artificial intelligence researcher at this point.

With all that said, the particular means by which we have achieved robotic self-recognition – namely, the comparison of dynamic Bayesian models – is interesting in that it is closely related to a different ability: action recognition. In computing the probability of the "animate other" model, the algorithm must necessarily compute the probability of each state of the "animate other." Though our particular implementation used only two states, "moving" or "not moving", the self-recognition algorithm would presumably work just as well, if not better, with more states in each model, which could correspond to either lower-level actions ("arm moving") or higher level ("running"), given an appropriate change in the representation of motion evidence. Indeed, the very computations that we have used as a subroutine to decide *whether* an entity is somebody else have previously been used to decide what that person is doing [4].

That the self-recognition decision runs a Bayesian model normally used for action recognition as a subroutine suggests an interesting hypothesis: that any species that can self-recognize, can also imitate. If animal self-recognition functions in a manner similar to our robot's, then the animal must necessarily produce a model of the other entity that includes a most likely hypothesis for what it is doing, prior to deciding whether the thing is itself or someone else. Presumably, the "animate other" model would need to be roughly analogous to the animal's self-model for the comparison to be accurate, meaning that the animal would compute the necessary hidden states required to produce the observed behavior. Having computed those states, the animal would then be able to enact those states itself. One might also tentatively suggest that "mirror neurons" that respond to specific actions regardless of their actor [19] would be tentatively implicated as providing some of the evidence nodes that are accessed by both the self model and the "animate other" model.

In our implementation, the conditional dependence structures of the "self" and "animate other" models are given to the robot to begin with, while these causal structures may actually be learned in humans and self-recognizing animals. In fact, the acquisition of these models, along with their relative complexity compared to our boolean-state models, may explain the relative disparity between animals and our robot in time to acquire self-recognition capabilities: humans take 2 years, not 4 minutes, to acquire mirror self-recognition. It is an interesting question as to whether learning the conditional dependence structures or learning the more complex kinematic states are more responsible for this disparity in time to learn. It is possible that humans possess the correct structures necessary to self-recognize from birth, but do not have sufficient data to give correct classifications until later in development. However, as roboticists, we cannot really defend the practicality of changing a system that takes 4 minutes to adapt to one that takes 2 years to adapt, and so a true model of human and animal self-recognition is a matter for another paper.

To the robot builder, we note that the method is quite easy to implement, robust, provides readily interpreted probabilities that can be used elsewhere for reasoning, and is fairly obviously extendable to more complicated models of the robot's self and the entities around it. Because a robot's appearance can change under various lighting conditions or in case of dirt or damage, a self-recognition algorithm that depends only on motion may have useful application beyond these more philosophical discussions.

A few caveats are in order, however. First, it is not clear how useful the "inanimate" model is to the decision process. Using background subtraction as our method for both segmentation and finding motion, it was often the case that unmoving objects registered as moving because of camera noise – particularly in the case of the lab's fluorescent lights, which produced constant lighting fluctuations too subtle for the human eye but easily picked out by our robot's cameras. In principle, if the "self" or "animate other" models included more states, and a more complicated relationship between those states and the evidence nodes, the "animate other" model would become less fit for modeling inanimate objects. It would then be more important to have a model expressly for the purpose of identifying inanimate objects. Here, however, it is not clear whether this category is doing much that is useful; its usefulness may only become apparent with a more sophisticated vision system or action model.

Second, it is easy to forget that the robot is here recognizing only its arm and hand as itself; background subtraction renders the rest of its body invisible. Ideally, a better vision system would segment the whole robot as a single object; the algorithm's decision of "self" based on the arm's motion would then apply to the whole robot. However, this might be wishing for a perfect technology that doesn't exist; after all, what *does* make our robot's head, arm, and central pole all part of the same object – but not the table on which the robot rests? Does it even make sense to count the central metal pole as a part of the robot, if it can't be actuated and has no sensors of its own? It is not clear to what extent such parts should be labeled as "self."

The limitations of background subtraction also currently restrict the robot's head from moving, making the robot unable to label its head as "self" in the mirror. If the robot possessed a representation of the environment with a fixed frame of reference, such that it could tell when it moved its head that one small thing was moving and not the whole environment, it would be straightforward for the robot to apply the motion models to the reflected head and discover it to be a part of itself. The same would be true of the body of a wheeled robot; being able to detect that its image in the mirror was moving while its own frame of reference shifted would be the only difficulty in getting such a robot to recognize its whole body as itself.

The researcher of robotic self-recognition must walk a fine line. On the one hand, any results must be strenuously underplayed, as the expectation in the popular mind that self-recognition is a huge dividing line between the sentient and the unaware is simply false, at least in the domain of artificial intelligence. On the other hand, the ability to reliably learn to recognize a new or changed body part is a useful application that requires some cleverness on the part of the designer, in the form of Bayesian reasoning over time, to quickly integrate all the information that the robot has accumulated over time while remaining robust to noise. Though our method is not a model of human or animal self-recognition, it seems plausible that perhaps some of the mechanisms we describe here – the symmetry between the "self" and "animate other" model, the bootstrapping of the probabilities (or neuronal weights) involved in each, and the efficient scaling of the beliefs based on current evidence through a forward model-like propagation – could have analogous mechanisms in biological entities that self-recognize.

References

[1] B. Amsterdam, Mirror self-image reactions before age two, Developmental Psychobiology 5 (4).

[2] L. E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Annals of Mathematical Statistics 41.

[3] P. Bloom, How Children Learn the Meanings of Words, MIT Press, Cambridge, Massachusetts, 2000.

[4] A. Dearden, Y. Demiris, Learning forward models for robots, in: Proc. IJCAI, Edinburgh, Scotland, 2005.

[5] G. Gallup, Self-awareness and the emergence of mind in primates, American Journal of Primatology 2 (1982) 237–248.

[6] G. G. Gallup, Chimpanzees: self-recognition, Science 167 (3914) (1970) 86–87.

[7] G. G. Gallup, J. Anderson, D. Shillito, The mirror test, in: M. Bekoff, C. Allen, G. Burghardt (eds.), The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition, MIT Press, 2002.

[8] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences, Basic Books, 1997.

[9] K. Gold, B. Scassellati, A Bayesian robot that distinguishes "self" from "other", in: Proceedings of the 29th Annual Meeting of the Cognitive Science Society, 2007.

[10] P. O. A. Haikonen, Reflections of consciousness: The mirror test, in: Proceedings of the 2007 AAAI Fall Symposium on Consciousness and Artificial Intelligence, available online: http://www.consciousness.it/CAI/CAI.htm, 2007.

[11] M. D. Hauser, C. T. Miller, K. Liu, R. Gupta, Cotton-top tamarins (*Saguinus oedipus*) fail to show mirror-guided self-exploration, American Journal of Primatology 53 (2001) 131–137.

[12] C. C. Kemp, A. Edsinger, What can I control?: The development of visual categories for a robot's body and the world that it influences, in: Proceedings of the International Conference on Development and Learning (Special Session on Autonomous Mental Development), Bloomington, IN, 2006.

[13] M. Lewis, M. W. Sullivan, C. Stanger, M. Weiss, Self development and self-conscious emotions, Child Development 60 (1989) 146–156.

[14] P. Michel, K. Gold, B. Scassellati, Motion-based self-recognition, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, 2004.

[15] S. Pinker, How the Mind Works, W. W. Norton and Company, 1997.

[16] J. M. Plotnik, F. B. M. de Waal, D. Reiss, Self-recognition in an asian elephant, PNAS 103 (45).

[17] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–296.

[18] D. Reiss, L. Marino, Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence, Proceedings of the National Academy of Sciences 98 (10).

[19] G. Rizzolatti, L. Fogassi, V. Gallese, Neurophysiological mechanisms underlying the understanding and imitation of action, Nature Reviews Neuroscience 2 (2001) 661–670.

[20] T. Suzuki, K. Inaba, J. Takeno, Conscious robot that distinguishes between self and others and implements imitation behavior, in: M. Ali, F. Esposito (eds.), Innovations in Applied Artificial Intelligence: 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Springer-Verlag, Bari, Italy, 2005.

[21] W. G. Walter, An imitation of life, Scientific American (1950) 42–45.

[22] Y. Yoshikawa, K. Hosoda, M. Asada, Cross-anchoring for binding tactile and visual sensations via unique association through self-perception, in: Proceedings of the Fourth International Conference on Learning and Development, San Diego, CA, 2004.

[23] Y. Yoshikawa, Y. Tsuji, K. Hosoda, M. Asada, 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, 2004.