Robotics and Autonomous Systems 57 (2009) 384–392

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Using probabilistic reasoning over time to self-recognize

Kevin Gold *, Brian Scassellati

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

ARTICLE INFO

Article history:
Received 15 May 2008
Accepted 30 July 2008
Available online 13 August 2008

Keywords:
Self-recognition
Robot
Mirror test
Dynamic Bayesian model
Animacy
Contingency

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 s before mirroring, the evidence was "remembered" across a full minute of mimicry.

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. 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 [13,6]. The use of a mirror to assay intelligence is an attractive idea, because the dividing line seems to so clearly segregate the intelligent species from the not-so-intelligent: among animal species, only apes [7,8], 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 [12]. 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 [14,11], 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 a 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 [20,11,21]. Because our method produces probabilities with clearly defined semantics, the results

* Corresponding address: Department of Computer Science, Wellesley College, 106 Central St., MA-02481 Wellesley, USA. Tel.: +1 714 931 9275; fax: +1 714 931 9275.

E-mail addresses: kevin.gold@yale.edu, kgold@wellesley.edu (K. Gold), scaz@cs.yale.edu (B. Scassellati).

0921-8890/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2008.07.006

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 share high mutual information with the robot's kinematic model [4]. 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 (Fig. 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 [10], while the experiments in Section 5, this introduction, and the conclusions are new to this paper.)

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.

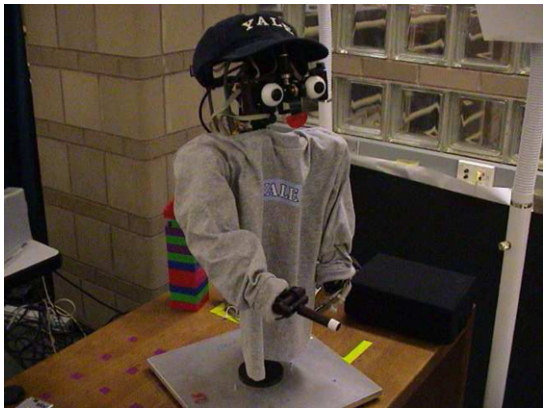

Fig. 1. Nico is an upper-torso humanoid robot with the arm and head kinematics of a one-year-old.

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.)

Fig. 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]

[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 evidence. Only the likelihood of the motion observations conditioned on motor activity is being calculated, not the joint likelihood of motor activity and motion. 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 [10].