BioSystems 128 (2015) 37–47

Contents lists available at ScienceDirect

BioSystems

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

Inference of other's internal neural models from active observation

Kyung-Joong Kim [a], Sung-Bae Cho [b],*

[a] Department of Computer Science and Engineering, Sejong University, Seoul, South Korea
[b] Department of Computer Science, Yonsei University, Seoul, South Korea

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form 6 January 2015
Accepted 19 January 2015
Available online 21 January 2015

Keywords:
Theory of mind
Robot
Physics-based simulation
Active learning
Neural network
Evolutionary computation
Estimation–exploration algorithm

ABSTRACT

Recently, there have been several attempts to replicate theory of mind, which explains how humans infer the mental states of other people using multiple sensory input, with artificial systems. One example of this is a robot that observes the behavior of other artificial systems and infers their internal models, mapping sensory inputs to the actuator's control signals. In this paper, we present the internal model as an artificial neural network, similar to biological systems. During inference, an observer can use an active incremental learning algorithm to guess an actor's internal neural model. This could significantly reduce the effort needed to guess other people's internal models. We apply an algorithm to the actor–observer robot scenarios with/without prior knowledge of the internal models. To validate our approach, we use a physics-based simulator with virtual robots. A series of experiments reveal that the observer robot can construct an "other's self-model", validating the possibility that a neural-based approach can be used as a platform for learning cognitive functions.

© 2015 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Robots can represent a simplified model of human behavior, whereby the robot senses its environment and reacts to various input signals. The robot's 'brain' controls its body in response to the input signals using artificial neural networks. The topology and weights of the neural network characterize the behavioral properties of the robot. Recently, several investigations have used robots in order to gain insight into human cognition by creating a simplified analogous problem (Bongard et al., 2006; Webb, 2001; Floreano and Keller, 2010). Bongard et al. built a starfish robot; however, it was unaware of its own body shape (Bongard et al., 2006). Using an estimation–exploration algorithm (EEA) (Bongard and Lipson, 2007), the robot was able to successfully create a self-model of its body shape using an iterative estimation and exploration procedure. In the estimation step, the robot searched multiple candidates to determine its body shape. Subsequently, in the exploration step, the algorithm determined the actions that most strongly agreed with the multiple candidate body shapes.

Unlike self-modeling, however, theory of mind (ToM) is a high-level cognitive function that models the mental states (beliefs, intents, desires, imagination, knowledge, *etc.*) of another entity. In robotic studies, robots have demonstrated the ability to mimic the behavior of humans or to decode the intentions of a third party (both human and robot). For example, Scassellati implemented Baron-Cohen's ToM model for the humanoid robot COG (Scassellati, 2002). Breazeal et al. demonstrated that an animal-like robot could pass the false-belief test widely used to test ToM in young children (Breazeal et al., 2005). Furthermore, Buchsbaum et al. carried out simulations in which one agent attempted to determine another agent's behavior using rat-like characters (Buchsbaum et al., 2005). In this particular study, the observer exploited his own behavior tree to infer others' intentions.

However, few reports have described the representation of another entity's mind as a neural circuit. Revealing an internal neural model based on observations is a challenging task. However, there is great potential for using neural networks as internal models, because it would mimic the underlying mechanisms of human representations in the form of neural connections. Many different definitions of the self and other's self-representations exist, ranging from symbolic states to complex neural models. For example, Bongard et al. (Bongard et al., 2006) used the morphological structure of a robot as a self-model. The robot had no physical model of itself on which to base an understanding, and attempted to construct models of its body using iterative estimation–exploration steps. Kim and Lipson used a simple feed-forward network to represent the minds of other (Kim and Lipson, 2009a,b).

In this paper, we propose the use of active incremental learning to infer the internal neural models of other entities both with and without prior knowledge (Fig. 1). We used two robots, referred to

* Corresponding author.
E-mail addresses: kimkj@sejong.ac.kr (K.-J. Kim), sbcho@cs.yonsei.ac.kr (S.-B. Cho).

http://dx.doi.org/10.1016/j.biosystems.2015.01.005
0303-2647/© 2015 Elsevier Ireland Ltd. All rights reserved.

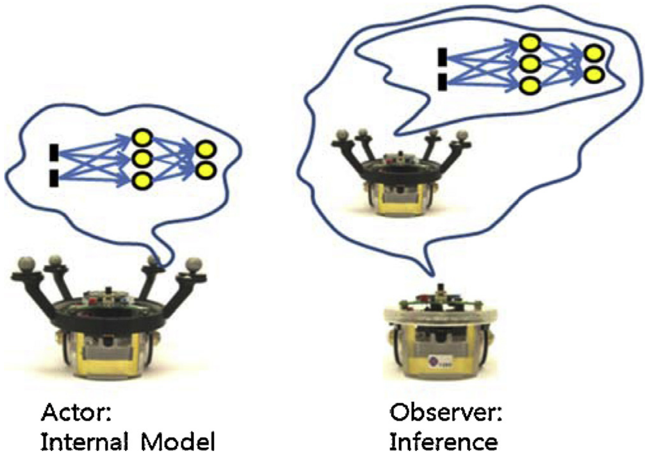

Fig. 1. Inference of other's internal models.

as the actor and the observer. The actor used a neural controller (implemented as an artificial neural network) to control its behavior based on sensory information. The observer monitored the behaviors of the actor and attempted to infer the actor's internal model from these observations. The observer used the inferred self-model of the actor to predict the actor's future behavior. In this approach, instead of programming the other's internal model manually, the observer attempted to predict the other's self-model interactively. The observer robot started from a single actor trajectory and invited the actor robot to demonstrate additional trajectories, which were then used to infer information about the actor's self-model using the EEA method (Bongard and Lipson, 2007).

In particular, we tested the impact that prior knowledge had on the actor's internal model. Initially, we assumed that the actor and observer were the same species and that the observer could use his self-model (neural topology). Therefore, the ToM problem is formulated as the inference of the connection weights given the shared structure. We subsequently assumed that the two robots are different species and that the actor could not use his self-model for the ToM. As a result, the observer needs to search for the architecture of the neural network and the weights simultaneously to infer the other's self-model. We used a physics-based simulation to run the ToM experiments, which show the potential of this approach given the two experimental conditions.

The rest of this paper is organized as follows. In Section 2 we describe related research, including the research on ToM in robots. In Section 3 we apply the estimation–exploration algorithm for the robotic ToM. Finally, in Section 4, we present our experimental results.

2. Background

2.1. Inference of other's mind in humans

ToM is the ability to attribute mental states to oneself and others, and to understand that others have different beliefs, desires, and intentions from one's own (Premack and Woodruff, 1978). The first paper on ToM, published in 1978 by Premack and Woodruff, posed the question, "Does the chimpanzee have a theory of mind"? Since then, many articles on ToM in human and non-human primates have been published (Call and Tomasello, 2008). Attempts have been made to reveal the existence of ToM in many species, including monkeys, and elephants, and ToM has been used to inform studies related to fundamental mechanisms and how certain conditions, such as autism, may develop (Baron-Cohen, 1995). Recently, brain imaging technology has been used to demonstrate the activation of specific areas in the brain associated with ToM (Siegal and Varley, 2002).

In the 30 years since the introduction of ToM, researchers have proven that chimpanzees have ToM, but that they cannot understand each other to the degree that humans do (Call and Tomasello, 2008). Herrmann et al. compared ToM ability among humans, chimpanzees, and orangutans using gaze-following and intention-understanding tasks (Herrmann et al., 2007) showing that humans outperformed chimpanzees and orangutans. In humans, ToM has been shown to be related to neural development disorders that are characterized by impaired social interaction and communication; for example, childhood autism may be associated with a deficit in ToM (Baron-Cohen, 1995). Baron-Cohen compared normal subjects and subjects with autism and Down syndrome using a belief question to test ToM, finding that subjects with Down syndrome were similar to the control group; however, 80% of autistic children failed to show ToM (Baron-Cohen, 1995).

How the ToM works is not well understood. There are several theories to explain these high-level cognitive functions. Several robotics researchers have used robots in attempts to better understand these theories (Scassellati, 2002; Breazeal et al., 2005). However, debate among neuroscientists concerning the evidence supporting these different hypotheses persists (Siegal and Varley, 2002; Saxe, 2009). ToM theories can be classified into three categories: modular, theory–theory, and executive function theories (Youmans, 2004).

- In the modular view (supported by Baron-Cohen, 1995), ToM is functionally dissociable from other cognitive functions, and it is assumed that there is one or more neural structures specifically dedicated to this function. Baron-Cohen assumed that the ToM process includes an intentionality detector, an eye-direction detector, a shared-attention mechanism, and a ToM mechanism (Baron-Cohen, 1995).
- According to the theory–theory school, a child has a theory about how other minds operate, which evolves over time and with experience.
- Some theorists argue that a distinct ToM does not exist and that executive functions are sufficient to explain the skills involved in ToM (Ozonoff et al., 1991).

We believe that developing and testing ToM models using robots may provide insight into some of these complex questions.

2.2. Computational approaches for ToM

Robots are increasingly being used as a platform to test theories of human behavior and cognition (Webb, 2001). In a broad sense, a robot includes virtual agents, characters, simulated robots, and real robots. Recently, interesting interdisciplinary research has shown the effectiveness of a robot-based approach. For example, Wischmann et al. investigated the emergence of communications using physics-based robot simulators (Wischmann et al., 2012). Bongard et al. also used robots and a physics-based robot simulator, based on the open dynamics engine, to demonstrate robot self-modeling (Bongard et al., 2006).

Because ToM is an important cognitive function in humans, researchers have applied the concept to virtual agents, virtual characters, simulated robots, and real robots (see Table 1 for further details). In most studies, more than two robots were used, and each assumed the role of either "actor" or "observer". The observer robots inferred the internal model of the actor robot from observations of the actor's behavior. If the observer was successful in revealing the internal model of the actor, the model used in the estimation could be used to predict subsequent behavior by the actor. Because this inference is a kind of reverse-engineering task,

Download English Version:

https://daneshyari.com/en/article/2075894

Download Persian Version:

https://daneshyari.com/article/2075894

Daneshyari.com