Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 28 (2014) 44–79

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Towards computational models of intention detection and intention prediction

Action editor: Paul Bello

Elisheva Bonchek-Dokow [a,*], Gal A. Kaminka [b,a]

[a] *Brain Science Research Center, Bar Ilan University, Israel*

[b] *MAVERICK Group, Computer Science Department, Bar Ilan University, Israel*

Available online 26 July 2013

Abstract

Intention recognition is one of the core components of mindreading, an important process in social cognition. Human beings, from age of 18 months, have been shown to be able to extrapolate intentions from observed actions, even when the performer failed at achieving the goal. Existing accounts of intention recognition emphasize the use of an intent (plan) library, which is matched against observed actions for recognition. These therefore cannot account for recognition of failed sequences of actions, nor novel actions. In this paper, we begin to tackle these open questions by examining computational models for components of human intention recognition, which emphasize the ability of humans to detect and identify intentions in a sequence of observed actions, based solely on the rationality of movement (its efficiency). We provide a high-level overview of intention recognition as a whole, and then elaborate on two components of the model, which we believe to be at its core, namely, those of intention detection and intention prediction. By *intention detection* we mean the ability to discern whether a sequence of actions has any underlying intention at all, or whether it was performed in an arbitrary manner with no goal in mind. By *intention prediction* we mean the ability to extend an incomplete sequence of actions to its most likely intended goal. We evaluate the model, and these two components, in context of existing literature, and in a number of experiments with more than 140 human subjects. For intention detection, our model was able to attribute high levels of intention to those traces perceived by humans as intentional, and vice versa. For intention prediction as well, our model performed in a way that closely matched that of humans. The work highlights the intimate relationship between the ability to generate plans, and the ability to recognize intentions.

Keywords: Intention recognition; Intention prediction; Cognitive modeling

1. Introduction

Intention recognition is one of the core processes of *mindreading*, an important component in social cognition. Intention recognition involves identifying the goal of an observed sequence of actions, performed by some acting agent. It is a process by which an agent can gain access to the goals of another, and predict its future actions and trajectories. While it is not sufficient, by itself, for full mental state attribution (e.g., it does not ascribe beliefs to the observed agent), it is of critical importance in social interaction, and is of obvious evolutionary benefit. Indeed, human beings, from age of 18 months, have been shown to be able to extrapolate intentions from observed actions, even when the performer failed at achieving the goal (Meltzoff, 1995).

Existing accounts of intention recognition in artificial intelligence (*plan recognition*) and machine vision (*activity recognition*) emphasize the use of an intent (plan, activity) library, which is matched against observed actions for recognition. These therefore cannot account for recognition of failed sequences actions, nor novel actions. Moreover, these accounts ignore cognitive science literature, which shows that the process involves a number of component

* Corresponding author.
E-mail addresses: le7.bonchek.dokow@gmail.com (E. Bonchek-Dokow), galk@cs.biu.ac.il (G.A. Kaminka).

1389-0417/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cogsys.2013.07.004

processes: recognizing the agent as capable of possessing intentions, recognizing that the observed action sequence is intentional, hypothesizing the intent of the sequence (even if the sequence results in a failing), and more (we discuss this in Section 2).

In this paper, we focus on modeling two of these components, *intention detection* and *intention prediction*. By *intention detection* we mean the ability to discern whether a sequence of actions has any underlying intention at all, or whether it was performed in an arbitrary manner with no goal in mind. By *intention prediction* we mean the ability to extend an incomplete sequence of actions to its most likely intended goal.

In particular, we focus on the use of rationality (efficiency) of an observed action trajectory or plan as a possible basis for intention recognition. We argue that while there are several different ways in which humans may carry out intention detection and prediction (discussed in detail in Section 2), it is often possible to determine a level of intentionality of an observed sequence of actions, based solely on the observed actions, and the ability to plan (optimally). We thus highlight a role for planning within recognition.

Following Section 2, which motivates our work in context of existing literature, we begin in Section 3 with a brief description of an abstract intention recognition model, in which two of the components are intention detection and intention prediction (we describe this abstract model to put the component processes in context of the larger area of research). We then provide a detailed account of the computational models underlying these processes from our perspective, which focuses on rationality. In Sections 4 and 5 we evaluate the hypothesized models for *intention detection* and *intention recognition* processes, respectively.

In particular, in Section 4, the intention detection model was evaluated in a discrete-state recreation of key experiments in humans (Meltzoff, 1995), and in detecting intentionality in activity recognition videos. In both settings for evaluating the first component of intention detection, the results confirm that our model closely matches human performance. Traces of action that were deemed by human observers as highly intentional, were ranked similarly by our model, while traces of action that were judged by humans as less intentional, achieved lower grades of intention by our model as well. Thus, the predictions of the model were successfully compared to those of human subjects. In addition, our findings show that our model proves useful for detecting sub-goals as well.

In a final set of experiments (Section 5), the intention prediction component was evaluated with data from human subjects, manipulating two-dimensional objects in a computer-based experiment. These experiments show equally promising results. Our model was able to predict the correct intention of various action traces with high accuracy, using our suggested heuristic. Two other heuristics are evaluated as well, and show significantly inferior prediction ability.

Finally, in Section 6, we discuss the significance of the results, highlighting several aspects such as the use of different measures of intentionality and the role of the intention detection method we propose, along-side and complementary to other methods in intention recognition. We also suggest possible directions for future research.

2. Background and related work

First introduced by Premack and Woodruff (1978), *theory of mind* (also: *folk psychology*, *mentalizing*, and *mind-reading*) is the ability to attribute mental states (beliefs, intents, desires, etc.) to oneself and to others. As originally defined, it enables one to understand that mental states can be the cause of others' behavior, thereby allowing one to explain and predict the observed actions produced by others. This ability enables a psychological attribution of causality to human acts, rather than the physical causality generally attributed to inanimate objects (Meltzoff, 1995).

Different accounts are given by psychologists for the mechanism underlying this ability. One of them, known as *simulation theory* (Gordon, 1986; Davies & Stone, 1995; Heal, 2003), has gained popularity and credibility lately, in part due to the discovery of mirror neurons (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Gallese & Goldman, 1998; Fogassi et al., 2005; Dapretto et al., 2005). In the words of Breazeal, Buchsbaum, Gray, Gatenby, and Blumberg (2005), simulation theory posits that by simulating another person's actions and the stimuli the other is experiencing using their own behavioral and stimulus processing mechanisms, humans can make predictions about the behaviors and mental states of the other based on the mental states and behaviors that they themselves would possess if they were in the other's situation. In short, by thinking "as if" we were the other person, we can use our own cognitive, behavioral, and motivational systems to understand what is going on in the head of the other.

Thus theory of mind is intimately related to imitation, in subtle ways (Meltzoff & Moore, 1992, 1994, 1995; Meltzoff & Decety, 2003; Meltzoff & Gopnik, 1993). On the one hand, basic imitation of movement is a precursor to the development of theory of mind skills, by laying the foundations for what Meltzoff calls the "like me" framework for recognizing and becoming an intentional agent (Meltzoff, 2007). Once the infant learns by imitation that her body, along with its inputs and outputs, is similar to those of the adults she sees around her, then she can simulate their behavior within her own mind. On the other hand, once this capacity is developed, theory of mind can be put to use for the explanation and prediction of actions observed.

This paper is motivated by one specific line of investigations on the relation between theory of mind and imitation, that begins with an experiment by Meltzoff (1995). The experiment makes use of infants' tendency to imitate, to explore their mindreading capabilities, and specifically