Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 19 (2006) 299–310
Neural Networks
www.elsevier.com/locate/neunet

2006 Special issue

A probabilistic model of gaze imitation and shared attention

Matthew W. Hoffman *, David B. Grimes, Aaron P. Shon, Rajesh P.N. Rao

Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA

Abstract

An important component of language acquisition and cognitive learning is gaze imitation. Infants as young as one year of age can follow the gaze of an adult to determine the object the adult is focusing on. The ability to follow gaze is a precursor to shared attention, wherein two or more agents simultaneously focus their attention on a single object in the environment. Shared attention is a necessary skill for many complex, natural forms of learning, including learning based on imitation. This paper presents a probabilistic model of gaze imitation and shared attention that is inspired by Meltzoff and Moore's AIM model for imitation in infants. Our model combines a probabilistic algorithm for estimating gaze vectors with bottom-up saliency maps of visual scenes to produce maximum a posteriori (MAP) estimates of objects being looked at by an observed instructor. We test our model using a robotic system involving a pan-tilt camera head and show that combining saliency maps with gaze estimates leads to greater accuracy than using gaze alone. We additionally show that the system can learn instructor-specific probability distributions over objects, leading to increasing gaze accuracy over successive interactions with the instructor. Our results provide further support for probabilistic models of imitation and suggest new ways of implementing robotic systems that can interact with humans over an extended period of time.

Keywords: Imitation learning; Shared attention; Gaze tracking; Human-robot interaction

1. Introduction

Imitation is a powerful mechanism for transferring knowledge from a skilled agent (the instructor) to an unskilled agent (or observer) using manipulation of the shared environment. It has been broadly researched, both in apes (Byrne & Russon, 1998; Visalberghy & Fragaszy, 1990) and children (Meltzoff & Moore, 1977, 1997), and in an increasingly diverse selection of machines (Fong, Nourbakhsh, & Dautenhahn, 2002; Lungarella & Metta, 2003). The reason for the interest in imitation in the robotics community is obvious: imitative robots offer rapid learning compared to traditional robots requiring laborious expert programming. Complex interactive systems that do not require extensive configuration by the user necessitate a general-purpose learning mechanism such as imitation. Imitative robots also offer testbeds for computational theories of social interaction, and provide modifiable agents for contingent interaction with humans in psychological experiments.

1.1. Imitation and shared attention

While determining a precise definition for 'imitation' is difficult, we find a recent set of essential criteria due to Meltzoff especially helpful (Meltzoff, 2005). An observer can be said to imitate an instructor when:

(1) The observer produces behavior similar to the instructor.
(2) The observer's action is caused by perception of the instructor.
(3) Generating the response depends on an equivalence between the observer's self-generated actions and the actions of the instructor.

Under this general set of criteria, several levels of imitative fidelity and metrics for imitative success are possible. Alissandrakis, Nehaniv, and Dautenhahn (2000, 2003) differentiate several levels of granularity in imitation, varying in the amount of fidelity the observer obeys in reproducing the instructor's actions. From greatest to least fidelity, the levels include:

(1) Path granularity: the observer attempts to faithfully reproduce the entire path of states visited by the instructor.
(2) Trajectory granularity: the observer identifies subgoals in the instructor's actions, and changes its trajectory over time to achieve those subgoals.

* Corresponding author.
E-mail addresses: mhoffman@cs.washington.edu (M.W. Hoffman), grimes@cs.washington.edu (D.B. Grimes), aaron@cs.washington.edu (A.P. Shon), rao@cs.washington.edu (R.P.N. Rao).

0893-6080/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.neunet.2006.02.008

(3) Goal granularity: the observer selects actions to achieve the same final goal state as the instructor (irrespective of the actual trajectory taken by the instructor).

Many of the imitation tasks that span the above levels of granularity require the instructor and observer to simultaneously attend to the same object or environmental state before or during imitation. Such simultaneous attention is referred to as shared attention in the psychological literature. Shared attention has even been found to exist in infants as young as 42 min old (Meltzoff & Moore, 1977). Yet, as with other human imitative behaviors, shared attention is a deceptively simple concept.

In seminal papers, Nehaniv and Dautenhahn (2000), and, separately, Breazeal and Scassellati (2001) proposed several complex questions that must be addressed by any robotic imitation learning system. Other groups (Jansen & Belpaeme, 2005; Billard, Epars, Calinon, Cheng, & Schaal, 2004) have applied a similar taxonomy to the design of imitative agents. Among these questions are two that directly relate to shared attention:

(1) How should a robot know what to imitate?
(2) How should a robot know when to imitate?

A system for shared attention must address exactly these questions. An imitative system must determine what to imitate; a system for shared attention must determine whether an instructor is present, and if so, which components of the instructor's behavior are relevant to imitation. In the scope of shared attention, this task encompasses both finding an instructor and the ability to recognize if no instructor is present.

Once an instructor has been located, the observer can turn to the question of where the instructor is directing his or her attention. This step combines the questions of what and when. The observer must first discern the instructor's focus using cues such as the instructor's gaze, body gestures, verbalization, etc. Determining what to imitate again comes into play as the observer must determine, which of these cues are being used to convey the instructor's intent. Further, for a fully autonomous system, the robot must be able to distinguish the intentionality of tasks—a head-shake differs greatly from a head-movement looking towards a specific object. The question of when to act is then raised: the observer must determine when it has acquired enough information to successfully imitate (cf. the exploration–exploitation trade-off in reinforcement learning).

Action can be taken once the observer has determined where to look, but the observer is now at an impasse: what really matters is the instructor's attentional focus. Consider, for example, a person told to look to the right. This information is not useful unless the person has knowledge about the current task or some method to determine why they must look right. Robotic observers learning from humans inevitably encounter the same obstacle: the robot can look right, but is unlikely to know the specific objects to which its attention is being directed. Further, for the observer to direct its search towards relevant objects or environment states, it must possess some method to segment the scene and identify relevant subparts. The observer must then be able to associate other factors with the scene, such as audio cues or task-dependent context, and identify the most salient segment. The pursuit of all-purpose imitation depends on having a model for saliency, i.e. a model of what components of the environmental state are important in a given task. Low-level saliency models can be generic, capturing image attributes such as contrast and color, but in this paper, we focus on more useful higher-level, task- or instructor-specific models, representing the observer's learned context-dependent knowledge of where to look.

Many different frameworks have been pursued for implementing biologically inspired imitation in robots. Broadly, frameworks can follow: (i) a developmental approach, where the robot builds a model of social behaviors based on repeated interactions with an instructor or caregiver (such as (Breazeal & Velasquez, 1998; Breazeal, Buchsbaum, Gray, Gatenby, & Blumberg, 2005; Calinon & Billard, 2005)); (ii) a biologically-motivated model, such as neural networks (Billard & Mataric, 2000) or motor models (Johnson & Demiris, 2005; Demiris & Khadhouri, 2005; Haruno, Wolpert, & Kawato, 2000); or (iii) a combination of development and brain modeling (Nagai, Hosoda, Morita, & Asada, 2003). Our model learns a model of perceptual saliency based on interaction with an instructor, bootstrapping the learned model using a neurally-inspired prior model for saliency (Itti, Koch, & Niebur, 1998), thus combining the developmental and modeling approaches.

As Nehaniv, Dautenhahn, Breazeal, and Scassellati note, the complex questions of what and when to imitate are just now being addressed by the robotics community. We do not claim to fully answer these questions, but we wish to draw a link between these questions with regard to imitation itself and the sub-task of shared attention. Previous robotic systems, such as those of Scassellati (1999), Demiris, Rougeaux, Hayes, Berthouze, and Kuniyoshi (1997), are able to track the gaze of a human instructor and mimic the motion of the instructor's head in either a vertical or horizontal direction. Richly contingent human–robot interaction comparable to infant imitation, however, has proven much more difficult to attain. Price (Price, 2003), for example, addresses the problem of learning a forward model of the environment (Jordan & Rumelhart, 1992) via imitation (see Section 3), although the correspondence with cognitive findings in humans is unclear. Other frameworks have been previously proposed for imitation learning in machines (Billard & Mataric, 2000; Breazeal, 1999; Scassellati, 1999), although without the probabilistic formalism being pursued in this paper. We view probabilistic algorithms as critical in cases like gaze tracking, where the instructor's gaze target is subject to a high degree of perceptual uncertainty. More recent imitation work has incorporated probabilistic techniques such as principal components analysis, independent components analysis, and hidden Markov models (Calinon & Billard, 2005; Calinon, Guenter, & Billard, 2005, 2006). This work has concentrated on using humanoid robots to imitate human motor trajectories, for example to write a character using a marker. We view our system as being complementary to these approaches: ideally, shared attention