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

could help humanoid robots to direct limited sensory and processing resources toward stimuli that are likely to allow enaction of future motor plans. Separately, Triesch and colleagues have used robotic platforms to study shared attention in infants (Fasel, Deak, Triesch, & Movellan, 2002), specifically examining the gaze imitation interaction between children and robots. For the purposes of this paper, we assume that the goal of the instructor is to direct the attention of the observer to an object of mutual interest. A unique contribution of our paper is the development of a probabilistic theory of shared attention; below we enumerate the benefits of such a model, show results from a robotic implementation of the model, and discuss the implications for neural and cognitive models of imitation.

This paper presents a Bayesian model that combines gaze imitation with saliency models to locate objects of mutual interest to the instructor and the observer. Bayesian models are attractive due to their ability to fuse multiple sources of information and handle noisy and incomplete data, all within a unifying mathematical formalism. The model described in this paper allows a robotic system to follow a human instructor's gaze to locate an object and over successive trials, learn an instructor- and task-specific saliency model for increased object location accuracy. Our biologically-inspired, model-based approach extends previous robotic gaze imitation results in three main ways: (i) it provides a Bayesian description of gaze imitation; (ii) it incorporates infant imitation findings into an algorithmic and model-based framework; and (iii) the system learns simple, context-dependent probabilistic models for saliency. Our results show the value of a Bayesian approach to developing shared attention between humans and robots.

Throughout this paper we use the term gaze imitation rather than head imitation to describe our process of gaze estimation—we do not, however, utilize eye-tracking in attaining this estimate. The model developed in this paper mirrors the learning apparatus utilized by young infants, specifically at the stage in their development where they are unable to distinguish between head movements and eye movements (Brooks & Meltzoff, 2002). The use of gaze imitation reinforces the idea that our goal is to track and imitate the instructor's gaze, whereas imitating the head movements is merely a byproduct of this process.

1.2. The active intermodal mapping model

At the highest level, our model is inspired by the work of Meltzoff and Moore, particularly their active intermodal mapping (AIM) hypothesis (Meltzoff & Moore, 1997). This hypothesis views infant imitation as a goal-directed, 'matching-to-target' process in which infants compare their own motor states (derived from proprioceptive feedback) with the observed states of an adult instructor. This comparison takes place by mapping both the internal proprioceptive states of the observer and the visual image of the instructor into a single, modality-independent space. Mismatch in this modality-independent space drives the motor planning system to perform corrective actions, bringing the infant's state in line with the adult's. In our case, the gaze angle of the instructor is extracted from the input image stream using an instructor-centric model, which allows gaze information to be easily converted to egocentric coordinates. Proprioceptive information from the robotic head provides information from encoders in the motors about current camera position, which can be compared with the target gaze angle for mismatch detection and motor correction. Fig. 1 juxtaposes the elements of AIM and our model. Other researchers have engaged in similar efforts to link infant development, specifically AIM, to systems for developmental robotics (Breazeal & Velasquez, 1998; Breazeal, 1999; Breazeal et al., 2005), although without the emphasis on probabilistic models.

Our present system models imitation in young infants. Infant studies have shown that, while younger infants use direction of adult head gaze to determine where to look, older infants use a combination of adult head direction and eye gaze (Brooks & Meltzoff, 2002). Other studies have shown that younger infants imitate adult gaze based on head movement (Moore, Angelopoulos, & Bennett, 1997; Lempers, 1979), while older infants can use static head pose (Lempers, 1979). As noted below, our system uses a Kalman filter of frame-by-frame observations to derive a robust estimate of gaze direction; thus, like younger infants, our system currently relies on observing head movement. More sophisticated feature detectors might allow inference of gaze direction from static images, although robust real-time gaze inference from single images remains an open problem in machine vision.

1.3. Motor models and Bayesian action selection

Many robotic systems model the environment, whether using a static map of an area or running a dynamic simulator of the world over time. Forward and inverse models (Jordan & Rumelhart, 1992) are commonplace in studies of low-level motor control. For example, Wolpert and colleagues have modeled paired forward and inverse models for motor control and imitation, and investigated possible neurological implementations (Blakemore, Goodbody, & Wolpert, 1998; Haruno et al., 2000). Forward and inverse models also provide a framework for using higher-level models of the environment to yield knowledge about actions to take, given a goal. Probabilistic forward models predict a distribution over future environmental states given a current state and an action taken from that state. Probabilistic inverse models encode a distribution over actions given a current state, desired next state, and goal state.

Learning an inverse model is the desired outcome for an imitative agent, since inverse models select an action given a current state, desired next state, and goal state. However, learning inverse models is difficult for a number of reasons, notably that environmental dynamics are not necessarily invertible; i.e. many actions could all conceivably lead to the same environmental state. In practice, it is often easier to acquire a forward model of environmental dynamics to make predictions about future state. By applying Bayes' rule, it becomes possible to rewrite a probabilistic inverse model in

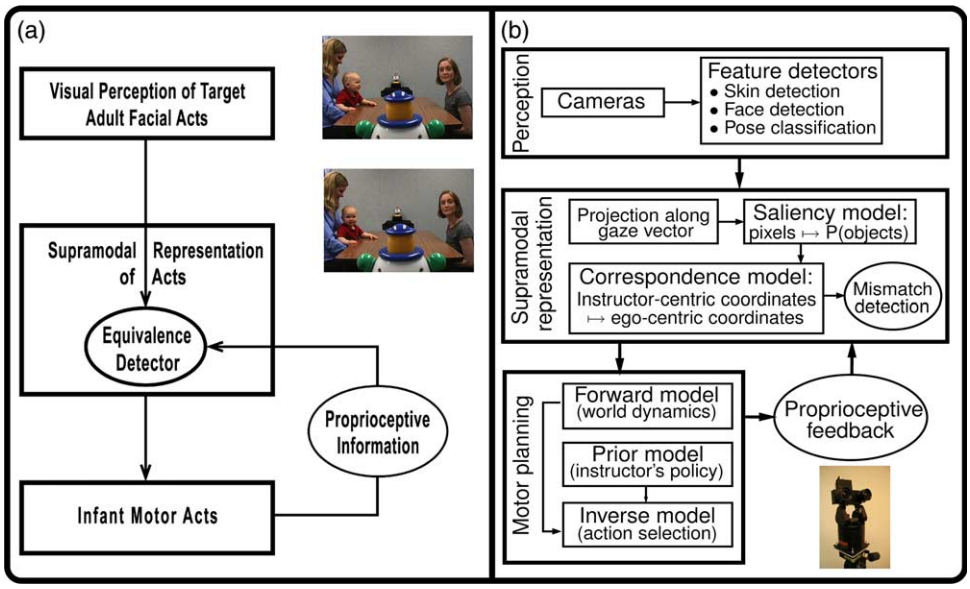

Fig. 1. Comparison between AIM and our model: (a) the active intermodal mapping (AIM) hypothesis of facial imitation by Meltzoff and Moore (1997) argues that infants match observations of adults with their own proprioception using a modality-independent representation of state. Mismatch detection between infant and adult states is performed in this modality-independent space. Infant motor acts cause proprioceptive feedback, closing the motor loop. The photographs show an infant tracking the gaze of an adult instructor (from (Brooks & Meltzoff, 2002)). (b) Our probabilistic framework matches the structure of AIM. Transforming instructor-centric coordinates to egocentric coordinates allows the system to remap the instructor's gaze vector into either a motor action that the stereo head can execute (for gaze tracking), or an environmental state (a distribution over objects the instructor could be watching) to learn instructor- or task-specific saliency.

terms of a forward model and a policy model (with normalization constant k) (Rao & Meltzoff, 2003; Rao et al., 2004)

$$P(a_t|s_t,s_{t+1},s_G) = kP(s_{t+1}|s_t,a_t)P(a_t|s_t,s_G) \quad (1)$$

where a_t is the action to be executed at time step t, s_t and s_{t+1} are states of the agent at time steps t and $t+1$, s_G is the desired goal state, and k is the normalization constant. Actions can be selected in one of two ways given such an inverse model. The observer can select the action with maximum posterior probability, or the observer can sample from $P(a_t|s_t,s_{t+1},s_G)$,strategy known as 'probability matching' (Krebs & Kacelnik, 1991), which seems to be used in at least some cases by the brain. Our present system uses only maximum a posteriori (MAP) estimates to select actions.

Previous robotic systems have employed the concepts of forward and inverse models for imitation (Demiris & Khadhouri, 2005; Haruno et al., 2000; Johnson & Demiris, 2005). Unlike these systems, which pair inverse and forward models for control and for prediction of sensory consequences, respectively, our system's inverse model is computed from a convolution of the forward and prior models as defined above. This Bayesian formulation simplifies the parameterization of the controller.

The present system does not learn a policy model, and instead assumes a uniform prior over actions that (according to the forward model) will move the imitator's motor state closer to the goal motor state. The system simply chooses the MAP estimate of a_t during training and testing based on observing the instructor's head pose. The policy model is implemented using a grid-based empirical distribution. Thus, according to the taxonomy given in Section 1, our model implements imitation at a goal-based level of granularity.

2. Probabilistic model of shared attention

In this section, we present a Bayesian approach to gaze imitation and shared attention, focusing on the interaction between one instructor and one observer (although this can readily be transformed in the case of multiple agents[1]). We accomplish this by presenting the observer with some set of objects with which the instructor will interact (e.g. by looking at one of the objects in each interaction). By watching the instructor at each time-step of this process, the observer is then able to learn a 'top-down saliency model' of these objects, encoding the instructor's object preferences.

Our framework provides a mathematically rigorous method for inferring the attentional focus of the instructor based on multiple environment cues. We draw a distinction between two sets of environmental cues, attentional (or instructor-based) and object-based. Attentional cues arise from observing the instructor's actions. Some examples of instructor-based cues are head gaze direction, saccadic eye movements, and hand gesture direction. Object-based cues are properties of the objects themselves: size, color, texture, sounds emitted, etc. This distinction allows us to view interaction in two stages: instructor-based cues give rough estimates for the focus of attention, whereas object-based information provides the ability to fine tune this estimate. Our specific use of instructor-based cues to provide an initial rough estimate is described in Section 3.

[1] For example by using a unique identifier for each agent such as cues provided by facial recognition. Separate saliency cues/preferences can be associated with each identifier.

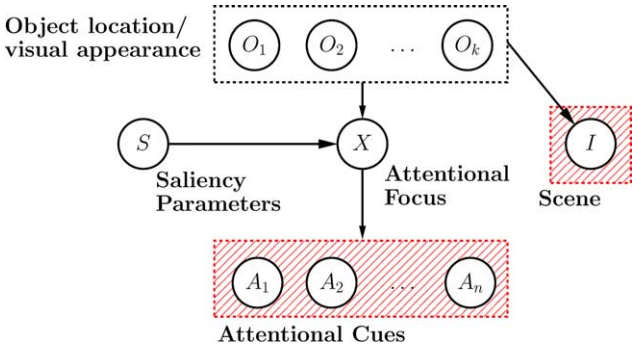

Fig. 2. Probabilistic formulation of shared attention. A Bayesian network describing the interaction between an instructor's focus of attention and perceived environmental cues. The model allows inference of the hidden attentional focus X based on the observed shaded variables. The set of variables $\{O_1,\ldots,O_k\}$ represent the location and appearance of each object. The observation I is an image of the scene containing objects, which may be attended to. The set of observations $\{A_1,\ldots,A_n\}$ represent attentional cues such as gaze or hand tracking. The variable S denotes the parameters of a task and instructor dependent model of saliency or object relevance.

Graphical models (specifically Bayesian networks) provide a convenient method for describing conditional dependencies such those between environmental cues and the attention of the instructor. Our graphical model used to infer shared attention is shown in Fig. 2. We denote the focus of the instructor by a random variable X. Depending on the specific application X can either represent a 3-dimensional real world location or a discrete object identifier, which is used in conjunction with a known map of object locations. For simplicity we first assume the latter case of a discrete object representation.

Object location and appearance properties are represented by the variables $O=\{O_1,\ldots,O_k\}$ for some k possible objects. The instructor's attentional focus X is modeled as being conditionally dependent on object location and visual properties. Thus, the instructor's top-down saliency or object relevance model is represented by the conditional probability $P(X|O)$. In general, this conditional probability model is task- and instructor-dependent. To account for this variability, we introduce the variable S, which parametrizes the top-down saliency model. This corresponds to the saliency model $P(X|O,S)$ as shown in Fig. 2. As an example of a top-down saliency model, suppose color is an important property of objects O_i. The variable S could then be used to indicate, for instance, how relevant a red object is to a particular task or instructor.

The attentional focus of the instructor is not directly observable. Thus, we model the attentional cues $\{A_1,\ldots,A_n\}$ as noisy observations of the instructor and their actions. Here, we consider n attentional observation models $P(A_i|X)$. In this paper, we utilize a probabilistic head gaze estimator as such an attentional observation model (see Section 3). However, it would be straightforward to incorporate additional information from observed gestures such as pointing.

The saliency model of the instructor (parameterized by S) is also considered unknown and not directly observable. Thus, we must learn S from experience based on interaction with the instructor. Initially, $P(S)$ is a uniform distribution, and thus $P(X|O,S)$ is equivalent to the marginal probability $P(X|O)$. An expectation maximization (EM) algorithm for incrementally learning S is described in Section 4.

A model of shared attention between a robot and a human instructor should be flexible and robust in unknown and novel environments. Thus, in this work we do not assume a priori knowledge about object locations and properties O_i nor the number of such objects k. Our model infers this information from an image of the scene I, as detailed in Section 4.

Ultimately, the goal of shared attention is to enable both the imitator and the instructor to attend to the same object. We select the object that the imitator attends to by computing the maximum a posteriori (MAP) value of X:

$$\overline{X} = \text{argmax}_X P(X|A_1,\ldots,A_n,I).$$

In order to infer the posterior distribution $P(X|A_1,\ldots,A_n,I)$ we first estimate MAP values of object locations and properties

$$\overline{O}_i = \text{argmax}_{O_i} P(I|O_i)P(O_i)$$

where $P(I|O_i)$ is determined using a low level saliency algorithm described in Section 4. The posterior can then be simplified using the Markov blanket of X, Blanket(X), i.e. the set of all nodes that are parents of X, children of X, or the parent of some child of X. Given this set of nodes the probability distribution $P(X|\text{Blanket}(X))$ is independent of all other nodes in the graphical model. Using the known information about objects present in the environment $\overline{O}_i$ we can calculate the probability distribution of X given its Markov blanket:

$$\begin{aligned} P(X|A_{1\ldots n},\overline{O}_{1\ldots k},S) &= P\big(X|\text{Blanket}(X)\big) \\ &= P\big(X|\text{Parents}(X)\big) \prod_{Z\in\text{Children}(X)} P\big(Z|\text{Parents}(Z)\big) \end{aligned} \tag{2}$$

$$P(X|A_{1\ldots n},\overline{O}_{1\ldots k},S) = P(X|S,\overline{O}_1,\ldots,\overline{O}_k)P(A_1|X)\cdots P(A_n|X). \tag{3}$$

3. Gaze following

A first step towards attaining shared attention is to estimate and imitate the gaze of an instructor. We use a probabilistic method proposed by Wu, Toyama, and Huang (2000), although other methods for head pose estimation may also be used. An ellipsoidal model of a human head is used to estimate pan and tilt angles relative to the camera. Inferred head angles are used in conjunction with head position to estimate an attentional gaze vector $g=A_i$ forming the attentional cue likelihood model $P(g|X)$.

The orientation of the head is estimated by computing the likelihood of filter outputs (within a bounding box of the head) given a particular head pose. During training, a filter output distribution is learned for each point on the three-dimensional mesh grid of the head. Thus, at each mesh point on the ellipsoid, filter responses for Gaussian and rotation-invariant Gabor at four different scales are stored. Our implementation of the Wu–Toyama method is able to estimate gaze direction in real-time (at 30 frames per second) on an average desktop computer.

The principal difficulty with this method is that it requires a tight bounding box around the head in testing and in training images for optimal performance. In both instances, we find the instructor's head using a feature-based object detection framework developed by Viola and Jones. This framework uses a learning algorithm based on the 'AdaBoost' algorithm to find efficient features and classifiers, and combines these classifiers in a cascade that can quickly discard unlikely features in test images. Features such as oriented edge and bar detectors are used that loosely approximate simple cell receptive fields in the visual cortex. We favor this method because of its high detection rate and speed for detecting faces: on a standard desktop computer, it can proceed at over 15 frames per second.

The face detection algorithm described above is only trained on frontal views of faces, allowing a narrow range of detectable head poses (plus or minus approximately 5–7° in pan and tilt). We circumvent this problem by first finding a frontal view of the face and then tracking the head across different movements using the Meanshift algorithm (Comaniciu, Ramesh, & Meer, 2000). The algorithm tracks non-rigid objects by finding the most likely bounding box at time t based on the distribution of color and previous positions of the bounding box. An attempt is made to minimize the movement in bounding box location between any two frames while also maintaining minimal changes in color between successive frames. The meanshift algorithm is used to track the position of the head over subsequent images, but this process does not always result in a tight bound. As a result, there is additional noise present in the head pose angle calculated using this bounding box. In order to account for this additional noise, a Kalman filter on the coordinates output by the meanshift tracker is utilized. This filtering of noisy gaze estimates based on an observed motion sequence is similar to the gaze imitation process in younger infants, who must observe head motion in order to follow the gaze of the instructor.

To summarize, the observer begins by tracking the instructor's gaze when the instructor looks at the observer, a traditional signal of attention. At this point the observer maintains the location of the instructor's head via a bounding box on the instructor's face as the instructor makes a head movement. A bounding box on the instructor's head allows the observer to determine the instructor's gaze angle at each point in this sequence using the previously learned ellipsoidal head model described earlier. The final gaze angle can then be determined from the observed head-motion sequence.

4. Estimating saliency

In humans, shared attention through gaze imitation allows more complex tasks to be bootstrapped, such as learning semantic associations between objects and their names, and imitating an instructor's actions on objects. Gaze imitation alone only provides a coarse estimate of the object that is the focus of the instructor's attention. Our model utilizes two other sources of information to fine tune this estimate: (1) bottom-up saliency values estimated from the prominent features present in the image (to facilitate object segmentation and identification), and (2) top-down saliency values encoding preferences for objects (S) learned from repeated interactions with an instructor.

Bottom-up saliency values for an image are computed based on a biologically-inspired attentional algorithm developed by Itti et al. (1998). This algorithm returns a saliency 'mask' (see Fig. 3(f)) where the grayscale intensity of a pixel is proportional to saliency as computed from feature detectors for intensity gradients, color, and edge orientation. The use of this algorithm allows interesting parts of the scene to be efficiently selected for higher level analysis using other cues. Such an approach is mirrored in the behavior and neuronal activity of the primate visual system (Itti et al., 1998). Thresholding the saliency mask and grouping similarly valued pixels in the thresholded image produces a set of discrete regions the system considers as candidates for objects. The ability to identify candidate objects is contingent on a sufficient separation placed between objects in the image. If two objects are located in positions such that they are within some small bound, or are overlapping, the algorithm will identify this region as one object. This is understandable, however, as distinguishing occluded objects would require some prior knowledge of the object appearance—which this low-level algorithm does not possess.

After repeated interactions with an instructor, the imitator can build a top-down context-specific saliency model of what each instructor considers salient—these instructor preferences are encoded in the prior probability over objects $P(X|S)$. As previously noted, this top-down model provides a method to reduce ambiguity in the instructor-based cues by weighting preferred objects more heavily. With no prior information, however, the distribution $P(X|S)$ is no different from $P(X)$.

We now consider a top-down saliency model, which is not domain dependent, yet enables the learning of task- and instructor-dependent information. We focus on leveraging generic object properties such as color (in YUV space) and size. Recall that top-down saliency corresponds to the conditional probability model $P(X|O,S)$. In our implementation, object appearance O_k is represented by a set of vectors $\mathbf{o}_i = \langle u_i, v_i, z_i \rangle$ where u_i and v_i are the UV values of pixel i, and where z_i is the size of the object (in pixels) from which this pixel is drawn.

As $\mathbf{o}$ is a continuous random variable, we utilize a Gaussian mixture model (GMM) to represent top-down saliency information. For each instructor, we need to learn a different Gaussian mixture model, thus it is intuitive to make S the parameters of a particular mixture model. Specifically, S represents the mean and covariance of C Gaussian mixture components, which are used to approximate the true top-down saliency model of the instructor.

Training the Gaussian mixture model is straightforward and uses the well known expectation maximization (EM) algorithm (Dempster et al., 1977). A set of data samples O_k from previous interaction is modeled as belonging to C clusters parameterized by S.

In inferring the attentional focus of the instructor, the system uses the learned model parameters S to estimate the prior (or

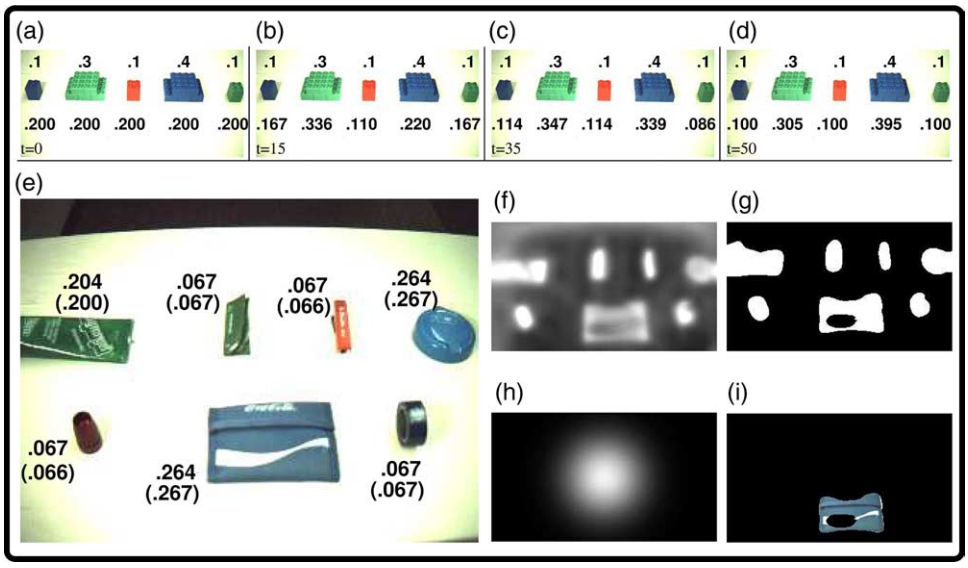

Fig. 3. Learning instructor-specific saliency priors: (a–d) the upper values give the true top-down saliency distribution. The lower values give the current estimate for this distribution, given t iterations. Progressing from (a–d) shows the estimate approaching the true distribution as number of iterations increases. (e) After training, we validate the learned saliency model using a set of testing objects. Next to each testing object is its estimated probability of saliency, with the true probability (according to the instructor) shown in parentheses. (f) A neurally-plausible bottom-up algorithm (Itti et al., 1998) provides a pixel-based, instructor-generic prior distribution over saliency, which the system thresholds to identify potentially salient objects. (g) Thresholded saliency map. (h) Intersection of instructor gaze vector and the table surface, with additive Gaussian noise. (i) Combination of (g) and (h) yields a MAP estimate for the most salient object in the test set (the blue wallet). (For interpretation of the reference to colour in this legend, the reader is referred to the web version of this article).

marginal w.r.t O) distribution over objects $P(X|S)$ for a specific instructor. The Gaussian mixture model yields an estimate on which object j the system should look at based on pixels in the connected components of the thresholded bottom-up saliency image. For each segmented object j in the scene the system first computes the maximum likelihood (ML) cluster label c_j for the object

$$c_j = \text{argmax}_{c\in C}\left(\left(\frac{1}{N_x}\sum_i^{N_x}\mathbf{o}_i - \mu_c\right)^{\mathrm{T}}\Sigma_c^{-1}\left(\frac{1}{N_x}\sum_i^{N_x}\mathbf{o}_i - \mu_c\right)\right), \tag{4}$$

where C is the set of Gaussian clusters in the mixture model and μ_c,Σ_c, respectively, denote the mean and covariance matrix for cluster c. The mixture model prior for Gaussian component c_j determines the a priori probability that the instructor will gaze at object j

$$P(X = j) = P(c_j), \tag{5}$$

where $P(c_j)$ is the probability that a point is drawn from the mixture component labeled c_j. The system finally combines this prior with the gaze and bottom-up saliency distributions to determine the MAP estimate of which object is being attended to. Fig. 3 illustrates the model in action.

4.1. Gaze imitation results

Our experimental set-up involved an instructor and a robotic observer (hereafter referred to as the robot) set at opposite ends of a table (shown in Fig. 4(a)). Initial tests focused on ascertaining the error in our gaze imitation algorithm. The model was first trained using video sequences from two different instructors looking in known directions. Once completed, the model was tested on in- and out-of-sample instructors gazing at two different positions on the table. Each different session was recorded as a success if the robot correctly aligned its gaze in the direction of the instructor's gaze. These tests showed accuracy of approximately 90%, both for in- and out-of-sample data (details are shown in Fig. 4(b)).

For these tests, the two points were at the center-line of the table approximately 1 m apart, while the robot and instructor were both approximately 0.5 m from the center of the table. In order to view these two positions, the instructor is required to gaze in approximately 45° in either direction. The main constraint on these distances results from the resolution of the robot's cameras: both robot and instructor must be close enough that the robot can discern individual objects when they are present on the table. As noted earlier, our system uses a low level saliency algorithm (Itti et al., 1998) to distinguish between objects, which limits the distance the robot can be from the table.

4.2. Incorporating learned instructor-specific priors

As illustrated by the example in Fig. 3(e), accurate gaze estimation does not alleviate the problems caused by a cluttered scene. Our next set of tests dealt with this problem of ambiguity. The instructor and robot are again positioned at a table as described earlier and objects are randomly arranged on the table; each pair of objects is separated by approximately 10 cm. The instructor is assumed to have a specific internal saliency model (unknown to the robot) encoding preferences for various objects. The instructor chooses objects based on this model. Once an object has been chosen, the instructor looks

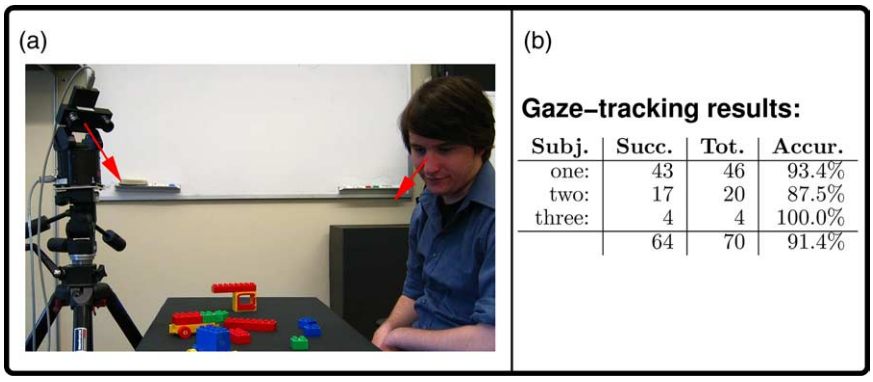

Fig. 4. Experimental setup and gaze tracking results: (a) the robotic observer tracks the instructor's gaze to objects on the table and attempts to identify the most salient object. (b) Accuracy of the gaze imitation algorithm in distinguishing between two locations, tested with three different subjects. Only the first of these subjects was in the training set.

towards the object, and the robot must track the instructor's gaze to the table in an attempt to determine the most salient object.

Once the robot has oriented to an object in the scene, we have the robot 'ask' the instructor whether it has correctly identified the instructor's object. We call a series of such attempts made by the robot to identify the instructor's object a trial. Monitoring the number of attempts made for each trial allows us to determine the accuracy of our system—as the number of trials increases, the robot should correctly identify objects with fewer and fewer attempts. A sequence of 20 successive trials was performed. Fig. 5 plots the accuracy of the combined gaze imitation and saliency model, where lower numbers represent more accurate object identification. Each sequence of trials was performed five times, with the values shown in Fig. 5 averaged over each sequence. The actual values plotted are the number of attempts made by the robot to identify the correct object, i.e. the number of incorrect proposals plus 1 for the last correct proposal.

For comparison, the first of these plots in Fig. 5, marked (a), shows the accuracy of the robot using random guesses to determine the object. The plot marked (b) uses gaze-tracking information only, and a random guess over objects in the robot's field of view. Finally, the plot in (c) combines the

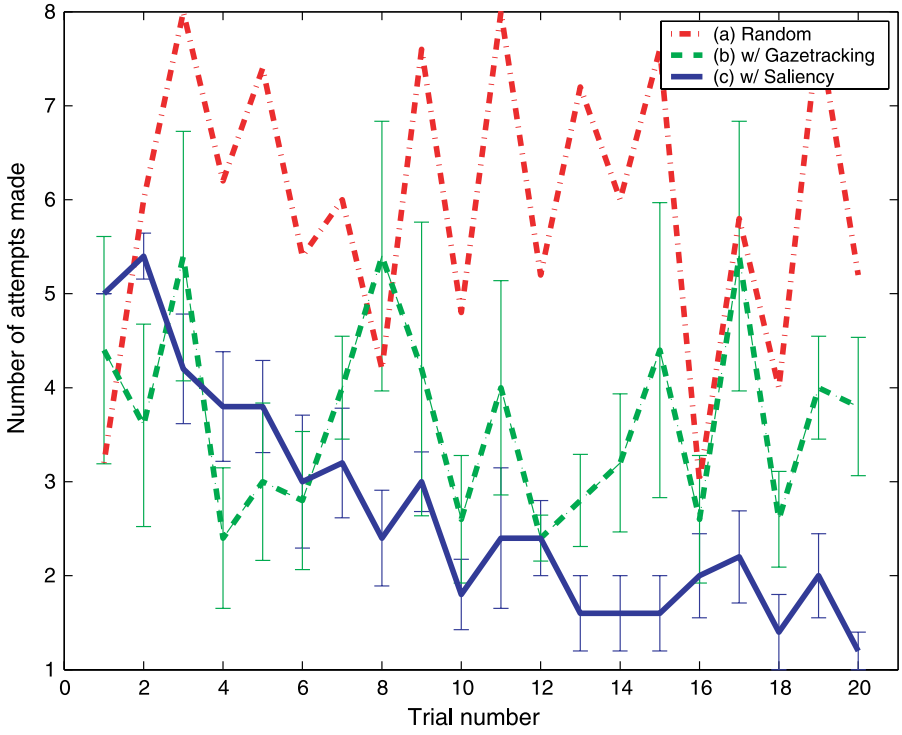

Fig. 5. Object localization accuracy over successive trials: the plot shows the accuracy of our system at locating 10 different objects to which the instructor is directing his attention, averaged over 5 sequences of trials. Values on the *y*-axis describe the average number of attempts made by the robot to identify the correct object, while values on the *x*-axis denote the trial number in the sequence. Line (a) shows the system using only random guesses to determine the object, while line (b) shows the inclusion of gaze information. Line (c) combines learned saliency information with gaze tracking, beginning with a uniform prior when no model is known. The error bars in this graph show the maximum and minimum number of attempts made for each trial.

information gained from gaze tracking and the current learned saliency model to propose the most likely object. It should be noted that the final two plots align closely for the first 5 to 6 steps, a trend which occurs as a result of how these trials were performed. The robot begins each trial with no prior information, as described in the previous section; as such it is expected that both this approach and just gaze-tracking perform with approximately the same accuracy. However, it can be seen that over time, the combined gaze imitation plus saliency model continues to improve, with steadily declining error, while the approach using gaze tracking remains at the same level of accuracy.

The saliency model (as seen in Fig. 3(a–d)) is not completely stable as of 20 trials; however, as we can see by viewing Fig. 3(a) and (c), the approximate likelihoods for each object class should be well established. One reason for ending these test sequences after 20 trials is that the robot is unlikely to perform with better accuracy than 2 or 3 attempts given the noise present in the system, and the possibility of the instructor gazing at rare (or less salient) objects. After 20 trials, tests using saliency information still perform much better than those using only gaze-imitation (an average of 2 versus 4 attempts at identification).

5. Relation to brain mechanisms of imitation

Our system does not explicitly model the neural architecture underlying imitative behaviors. In many cases, the neuroscience of imitation remains unclear. We can nonetheless draw analogies between components of our system and brain areas hypothesized as important for imitation. For example, the feature detectors used in the prior saliency algorithm (Itti et al., 1998) are based on center-surround detectors for image intensity and for each color channel, designed around the well-known properties of bipolar cells in retinal ganglia. Orientation detectors in the algorithm are based on simple cells of visual cortex area V1.

Other vision-based components of our algorithm have straightforward analogues in terms of brain structure. For example, the detection algorithm used by our system to localize the instructor's face (Viola & Jones) employs numerous Haar wavelet-like filters, measuring quantities such as the contrast between eyes and the bridge of the nose, etc. Very similar neuronal responses have been noted in primate inferotemporal cortex (IT) during face recognition tasks (Yamane, Kaji, & Kawano, 1988). Identification of facial features invariant to the relative viewpoint of the observer and instructor is vital to the success of our system, and is also a feature found in IT neurons (Booth & Rolls, 1998).

Although a full review of the possible neurological mechanisms underpinning AIM is beyond the scope of this paper, we note the pivotal discovery of 'mirror neurons' (Buccino et al., 2001; di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolati, 1992; Rizzolatti & Craighero, 2004; Rizzolatti, Fogassi, & Gallese, 2000), initially discovered in the macaque ventral premotor area F5, and later found in posterior parietal cortex and elsewhere. Mirror neurons fire preferentially when animals perform hand-related, manipulative motor acts, as well as when animals observe other agents (humans or monkeys) perform similar acts. Recent event-related fMRI results (Johnson-Frey et al., 2003) suggest that the left inferior frontal gyrus performs a similar function in humans, and that this area responds primarily to images of a goal state rather than to observations of a particular motor trajectory. Mirror neurons provide a plausible mechanism for the modality-independent representation of stimuli hypothesized by AIM.

The 'motor planning' aspects of our system shown in Fig. 1(b) are also linked to recent psychophysical and neurological findings. A critical component of our system is the predictive, probabilistic forward model that maps a current state and action to a distribution over future states of the environment. Imaging and modeling studies have implicated the cerebellum in computing mismatch between predicted and observed sensory consequences (Blakemore et al., 1998; Blakemore, Frith, & Wolpert, 2001; Haruno et al., 2000). Furthermore, recent papers have examined the potentially critical importance of information flow between cerebellum and area F5 during observation and replay of imitated behaviors (Iacoboni, 2005; Miall, 2003).

Based on our experimental results we make predictions about the reaction time of a human observer in obtaining shared attention with the instructor. We define reaction time as the time required for the subject to attend to a target object after observing the instructor's gaze. Reaction time can be predicted by combining experimental error rates during saliency learning (shown in Fig. 5) and a model of human eye movement (Carpenter, 1988). The experimental scenario we consider consists of a table 1 m from the observer with ten uniformly scattered objects. Saccade duration is modeled as linearly dependent on the angular distance between the various objects. We assume a mean saccade delay of 200 ms, which is consistent with such medium amplitude saccades (Carpenter, 1988). Fig. 6 shows the predicted reaction time and demonstrates how reaction time exponentially decreases as the observer learns the (non-uniform) preferences encoded by the instructor's object saliency distribution.

Based on the results in Section 4, our model makes the following psychophysical predictions for gaze following between an observer and an instructor in a cluttered environment, when the observer is initially ignorant of the instructor's saliency preferences:

- In the absence of previous experience with the instructor, observers will preferentially attend to objects within a region consistent with the observed gaze, and within that region to objects with high prior salience: regions of high contrast or high-frequency texture.
- The observer's error rate (percentage of objects incorrectly fixated by our system, or reaction time in the case of human infants) will decline exponentially in the number of trials (see Figs. 5 and 6) as the observer learns the preference of the human instructor.

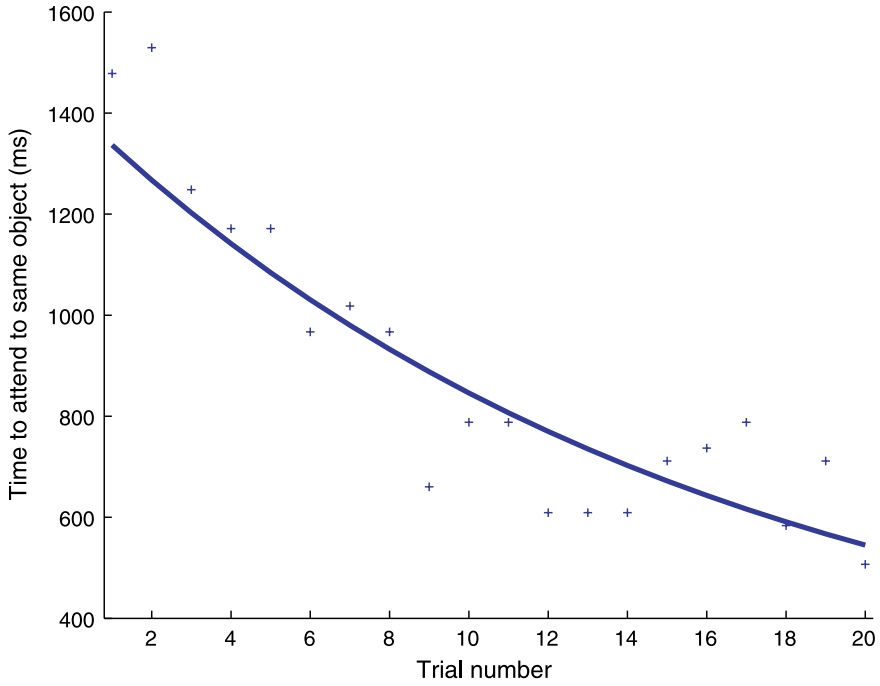

Fig. 6. Predicted time of obtaining shared attention during learning: the predicted reaction time of the observer after observing the instructor's gaze is plotted against the trial number. Reaction time is computed using a saccade duration model based on saccade latency and amplitude. Note that after each trial the observer better learns the instructor's (non-uniform) object saliency distribution. We plot an exponential curve fitted to the experimental data to illustrate the overall effect of saliency learning combined with gaze following behavior.

6. Conclusion

Gaze imitation is an important prerequisite for a number of tasks involving shared attention, including language acquisition and imitation of actions on objects. We have proposed a probabilistic model for gaze imitation that relies on three major computational elements: (1) a model-based algorithm for estimating an instructor's gaze angle, (2) a bottom-up image saliency algorithm for highlighting 'interesting' regions in the image, and (3) a top-down saliency map that biases the imitator to specific object preferences of the instructor as learned over time. Probabilistic information from these three sources are integrated in a Bayesian manner to produce a maximum a posteriori (MAP) estimate of the object currently being focused on by the instructor. We illustrated the performance of our model using a robotic pan-tilt camera head and showed that a model that combines gaze imitation with learned saliency cues can outperform a model that relies on gaze information alone.

The model proposed in this paper is closely related to the model suggested by Breazeal and Scassellati (2001). They too use saliency, both determined by an object's inherent properties (texture, color, etc) and by task context, to determine what to imitate in a scene, and use prior knowledge about social interactions to recognize failures and assist in fine-tuning their model of saliency. A similar system is put to further use with Kismet (Breazeal & Velasquez, 1998) (and more recently with Leonardo (Breazeal et al., 2005)). Breazeal and Scassellati's results are impressive and their work has been important in illustrating the issues that must be addressed to achieve robotic imitation learning. Our model differs from theirs in its emphasis on a unifying probabilistic formalism at all levels.

The early work of Demiris et al. (1997) on head imitation demonstrated how a robotic system can mimic an instructor's head movements. The system, however, did not have a capacity for shared attention in that the system made no attempt to follow gaze and find objects of interest. The work of Nagai et al. (2003) more closely investigates joint attention in robotic systems, focusing on the use of neural networks to learn a mapping between the instructor's face and gaze direction. Since, it relies on neural networks, their model suffers from many of the shortcomings of neural networks (e.g. ad hoc setting of parameters, lack of easy interpretation of results, etc.) that are avoided by a rigorous probabilistic framework.

The importance of gaze imitation has been argued throughout this paper, but we view gaze imitation as a building block towards the much-more important state of shared attention. In attaining a full-fledged shared attention model, we foresee the use of many different attentional and saliency cues. Such varied cues could be integrated into a graphical model similar to that shown in Fig. 2. One important attentional cue would include the hands of the instructor, or 'grasping motions' while interacting with objects. Fast, robust identification of hands and hand-pose is still an open problem in machine vision, one of the reasons why this important cue was not used in this paper.

In the future, we hope to extend our model to more complicated and varied saliency cues, as well as integrating more complex attentional cues. Specifically, such a system

would allow more sophisticated forms of human–robot interactions using a humanoid robot. Our algorithmic framework is hardware-agnostic, except for the forward model; the algorithm for instructor head pose estimation and the instructor-specific prior model will not change under this platform. Once we learn the forward dynamics of the humanoid's head, gaze imitation and saliency model learning will employ the same code base as the one for the Biclops head used in this paper. This extension could in turn enable more complex imitative tasks to be learned such as building objects from Lego™ blocks from demonstration. A learned instructor- or task-specific saliency model would bias the selection of Lego™ blocks of a particular color or shape during the construction of an object. We also anticipate extending our probabilistic model to accommodate more instructor-based cues (such as auditory information and pointing) to further increase gaze targeting accuracy.

References

Alissandrakis, A., Nehaniv, C. L., & Dautenhahn, K. (2000). Learning how to do things with imitation. In *Proceeding 'learning how to do things' (AAAI fall symposium series)*, (pp. 1–8).

Alissandrakis, A., Nehaniv, C. L., & Dautenhahn, K. (2003). Solving the correspondence problem between dissimilarly embodied robotic arms using the ALICE imitation mechanism. In *Proceedings of the second international symposium on imitation in animals & artifacts* (pp. 79–92).

Billard, A., Epars, Y., Calinon, S., Cheng, G., & Schaal S. (2004). Discovering optimal imitation strategies. *Robotics and Autonomous Systems*, 47, 2–3.

Billard, A., & Mataric, M. J. (2000). A biologically inspired robotic model for learning by imitation. In C. Sierra, M. Gini, & J. S. Rosenschein (Eds.), *Proceedings of the fourth international conference on autonomous agents* (pp. 373–380). Barcelona, Catalonia, Spain: ACM Press.

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (2001). The cerebellum is involved in predicting the sensory consequences of action. *Neuroreport*, *12*, 1879–1884.

Blakemore, S. J., Goodbody, S. J., & Wolpert, D. M. (1998). Predicting the consequences of our own actions: The role of sensorimotor context estimation. *Journal of Neuroscience*, *18*(18), 7511–7518.

Booth, M. C. A., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, *8*(6), 510–523.

Breazeal, C. (1999). Imitation as social exchange between humans and robots. In *Proceedings of the artificial intelligence and the simulation of behaviour* (pp. 96–104).

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., & Blumberg, B. (2005). Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life, 11*, 31–62.

Breazeal, C., & Scassellati, B. (2001). Challenges in building robots that imitate people. In K. Dautenhahn, & C. Nehaniv (Eds.), *Imitation in animals and artifacts*. Cambridge, MA: MIT Press.

Breazeal, C., & Velasquez, J. (1998). Toward teaching a robot 'infant' using emotive communication acts. In *Proceedings of the 1998 simulation of adaptive behavior, workshop on socially situated intelligence* (pp. 25–40).

Brooks, R., & Meltzoff, A. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, *38*, 958–966.

Buccino, G., Binofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience*, *13*, 400–404.

Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences*, *21*, 667–721.

Calinon, S., & Billard, A. (2005). Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM. In *Proceedings of the international conference on machine learning (ICML)*.

Calinon, S., Guenter, F., & Billard, A. (2005). Goal-directed imitation in a humanoid robot. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.

Calinon, S., Guenter, F., & Billard, A. (2006). On learning the statistical representation of a task and generalizing it to various contexts. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.

Carpenter, R. H. S. (1988). *Movements of the eyes* (2nd ed.). London: Pion Ltd.

Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 142–151).

Demiris, Y., & Khadhouri, B. (2005). Hierarchical, attentive multiple models for execution and recognition (HAMMER). In *Proceedings of the ICRA workshop on robot programming by demonstration*.

Demiris, J., Rougeaux, S., Hayes, G., Berthouze, L., & Kuniyoshi, Y. (1997). Deferred imitation of human head movements by an active stereo vision head. In *Proceedings of the sixth IEEE international workshop on robot human communication*.

Dempster, A.P., Laird, N.M., & Rubin, D. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolati, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*, 176–180.

Fasel, I., Deak, G. O., Triesch, J., & Movellan, J. R. (2002). Combining embodied models and empirical research for understanding the development of shared attention. In *Proceeding of ICDL 2*.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2002). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 142–166.

Haruno, M., Wolpert, D., & Kawato, M. (2000). MOSAIC model for sensorimotor learning and control. *Neural Computation*, *13*, 2201–2222.

Iacoboni, M. (2005). Understanding others: Imitation, language, empathy. In S. Hurley, & N. Chater (Eds.), *Perspectives on imitation: From mirror neurons to memes. Mechanisms of imitation and imitation in animals: Vol. 1*. Cambrdige, MA: MIT Press.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jansen, B., & Belpaeme, T. (2005). Goal-directed imitation through repeated trial-and-error interactions between agents. In *Proceedings of the workshop on social mechanisms of robot programming by demonstration*.

Johnson, M., & Demiris, Y. (2005). Hierarchies of coupled inverse and forward models for abstraction in robot planning, recognition and imitation. In: *Proceedings of the AISB symposium on imitation in animals and artifacts* (pp. 69–76).

Johnson-Frey, S. H., Maloof, F. R., Newman-Norlund, R., Farrer, C., Inati, S., & Grafton, S. T. (2003). Actions or hand–objects interactions? Human inferior frontal cortex and action observation. *Neuron*, *39*, 1053–1058.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*, 307–354.

Krebs, J. R., & Kacelnik, A. (1991). Decision making. In J. R. Krebs, & N. B. Davies (Eds.), *Behavioural ecology*. 3rd ed. (pp. 105–137). Oxford: Blackwell Scientific Publishers.

Lempers, J. D. (1979). Young children's production and comprehension of nonverbal deictic behaviors. *Journal of Genetic Psychology*, *135*, 93–102.

Lungarella, M., & Metta, G. (2003). Beyond gazing, pointing, and reaching: A survey of developmental robotics. In *EPIROB '03*, pp. 81–89.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*, 75–78.

Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting*, *6*, 179–192.

Meltzoff, A. N. (2005). Imitation and other minds: The 'like me' hypothesis. In S. Hurley, & N. Chater (Eds.), *Perspectives on imitation: From cognitive neuroscience to social science* (pp. 55–77). Cambridge, MA: MIT Press.

Miall, R. C. (2003). Connecting mirror neurons and forward models. *Neuroreport*, *14*, 2135–2137.

Moore, C., Angelopoulos, M., & Bennett, P. (1997). The role of movement in the development of joint visual attention. *Infant Behavior and Development*, *20*(1), 83–92.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). Emergence of joint attention based on visual attention and self learning. In *Proceedings of the second international symposium on adaptive motion of animals and machines*.

Nehaniv, C. L., & Dautenhahn, K. (2000). Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications. In J. Demiris, & A. Birk (Eds.), *Interdisciplinary approaches to robot learning*. Singapore: World Scientific Press.

Price, B. (2003). *Accelerating reinforcement learning with imitation*. PhD thesis, University of British Columbia.

Rao, R. P. N., & Meltzoff, A. N. (2003). Imitation learning in infants and robots: Towards probabilistic computational models. In *Proceedings of the artificial intelligence and the simulation of behaviour*.

Rao, R. P. N., Shon, A. P., & Meltzoff, A. N. (2004). A Bayesian model of imitation in infants and robots. In K. Dautenhahn, & C. Nehaniv (Eds.), *Imitation and Social Learning in Robots, Humans, and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge: Cambridge University Press.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2000). Mirror neurons: Intentionality detectors? *International Journal of Psychology*, *35*, 205.

Scassellati, B. (1999). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Lecture Notes in Computer Science*, *1562*, 176–195.

Viola, P., & Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*.

Visalberghy, E., & Fragaszy, D. (1990). *Do monkeys ape? In Language and intelligence in monkeys and apes: Comparative developmental perspectives*, Cambridge University Press (pp. 247–273).

Wu, Y., Toyama, K., & Huang, T. (2000). Wide-range, person- and illumination-insensitive head orientation estimation. In *AFGR00* (pp. 183–188).

Yamane, S., Kaji, S., & Kawano, K. (1988). What facial features activate face neurons in the inferotemporal cortex of the monkey? *Experimental Brain Research*, *73*(1), 209–214.