A Bayesian Active Vision Architecture for Shared Attention

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

CSE Department, Box 352350; University of Washington
Seattle WA 98195 USA

Abstract

Shared attention refers to the simultaneous perceptual focus of two or more agents on a single object in their shared environment. Shared attention is an important tool in building robotic systems capable of complex, natural forms of learning, such as imitation. This paper presents a probabilistic framework, based on Meltzoff and Moore's AIM model for imitation in infants, that emulates infants' capacity for shared attention. We also show the importance of context-dependent saliency for disambiguating visual elements, and bootstrapping more complex interactions between humans and robots. Our results demonstrate the value of a Bayesian perspective on shared attention and its application to imitation.

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 [9,28] and children [19,20], and in an increasingly diverse selection of machines [13,18]. The attraction of imitation for robotics is obvious: imitative robots offer drastically reduced costs compared to robots requiring expert programming. Complex interactive systems that do not require extensive configuration by the user necessitate the use of a general-purpose learning mechanism such as imitation.

Email addresses: mhoffman@cs.washington.edu (Matthew W. Hoffman), aaron@cs.washington.edu (Aaron P. Shon), grimes@cs.washington.edu (David B. Grimes), rao@cs.washington.edu (Rajesh P.N. Rao).

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 Overview of shared attention

Successful imitation requires that instructor and observer simultaneously attend to the same object or environmental state. Such simultaneous attention is referred to as *shared attention* in the psychological literature. Shared attention between a human instructor and observer is often taken for granted, and has been found to exist in infants as young as 42 minutes old [19]. Yet, as with other human behaviors, shared attention is a deceptively simple concept, overlaying many difficult problems.

In a recent paper, Breazeal and Scassellati proposed several complex questions that must be addressed by any robotic imitation learning system [5]. 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 its 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, verbalizations, 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. The question of *when* to act is then raised: the observer must determine when it has acquired enough information to successfully imitate. This is closely related to the exploration-exploitation tradeoff seen 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 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. Ideally, saliency models would be task- or instructor-specific, representing the observer's learned context-dependent knowledge of how to allocate attentional resources.

Robotic systems, such as those of Scassellati and Demiris [25,11], are already able to demonstrate impressive mimicry results. Both groups have separately focused on tracking the gaze of a human instructor and mimicking 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 [22], for example, addresses the problem of learning a forward model of the environment [16] via imitation (see section 1.3), although the correspondence with cognitive findings in humans is unclear. Other frameworks have been previously proposed for imitation learning in machines [1,3,25], but many of these are not designed around a coherent probabilistic formalism. Many of these approaches seem grounded in deterministic responses to environmental states—an approach which fails to capture the variability and error inherent in sensory input. Separately, Triesch and colleagues have used robotic platforms to study shared attention in infants [12], specifically examining the gaze following interaction between children and robots.

In this paper, we present a Bayesian framework that combines bottom-up attentional tracking with top-down saliency models to identify objects in a scene. Our algorithms employ Bayesian inference because of its robustness to noise and missing data, tractability under large data sets, and unifying mathematical formalism. The robotic system described in this paper tracks a human instructor's gaze gestures to locate an object, then learns an instructor and task-specific saliency model. Our biologically-inspired, model-based approach extends previous robotic gaze imitation results in three main ways: i) it provides a Bayesian description of imitation in general, and gaze tracking in particular; ii) it incorporates infant imitation findings into a rigorous 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.

1.2 The Active Intermodal Mapping model

At the highest level, our model of human-robot interaction is based on the work of Meltzoff and Moore, particularly the Active Intermodal Mapping (AIM) hypothesis [20]. 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. Fig. 1 juxtaposes the elements of AIM and our system.

Other researchers have engaged in similar efforts to link infant development, specifically AIM, to systems for developmental robotics [6,4,3].

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 [16] 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 [2,14]. 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 terms of a forward model and a policy model (with normalization constant k) [23,24]:

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

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 $P(a_t|s_t, s_{t+1}, s_G)$, a strategy known as "probability matching" [17], which seems to be used in at least some cases by the brain. Our present system uses only maximum *a priori* (MAP) estimates to select actions, which suffices to demonstrate the value of our approach.

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 Biclops' 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.

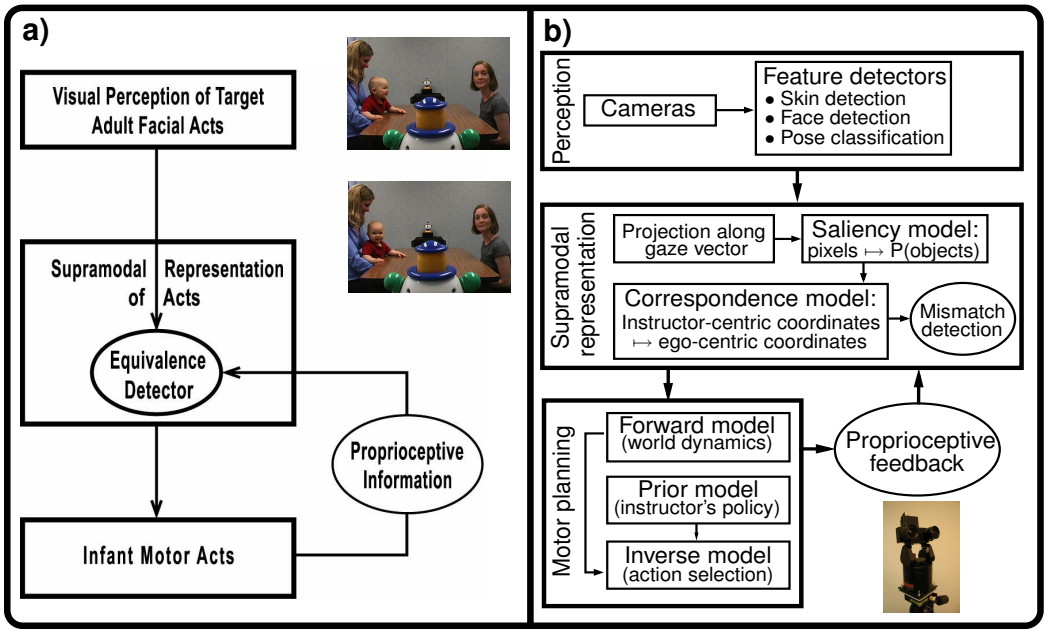


Fig. 1. **Active Intermodal Mapping hypothesis for infant imitation:** (a) The Active Intermodal Mapping (AIM) hypothesis of facial imitation by Meltzoff and Moore [20] argues that infants match observations of adults with their own proprioception using a modality-independent representation of state. 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 [7]). (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.

2 Probabilistic framework

The power of our approach to imitation lies in our probabilistic model of shared attention. We have argued why shared attention is important for building robots that can imitate, but just as important is a model that is strongly grounded in a rigorous probabilistic formalism. Ad hoc algorithms are not satisfying solutions in that they provide no underlying intuition for the imitative process. Nor are these approaches readily extensible or reviewable by the tools of mathematics.

In this paper we present a Bayesian approach to shared attention, focusing on the interaction between one instructor and one observer (although this can readily be transformed in the case of multiple agents). We accomplish this by presenting the observer with some set of objects with which the instructor will interact. By watching the instructor at each time-step of this process, the observer is then able to learn the saliency model of these objects, which it can then apply to further interactions.

In this framework we provide a distinction between two sets of environmental cues, attentional and saliency. Attentional cues provide information about the instructor and the direction of their attention—it is easiest to think of these as tracking the instructor's body: gaze tracking, hand tracking, etc. Saliency cues are best thought of as identifiers attached to the objects themselves: size, color, texture, audio cues, etc. This distinction allows us to view interaction in two stages: the attentional cues give rough estimates for the focus of attention, whereas the saliency cues provide the ability to fine tune this focus. The use of attentional cues to provide an initial rough estimate is detailed in section 3.

Let X be a discrete random variable representing the object to which the instructor is attending. The observer can combine a learned saliency model S with other attentional cues in the scene $\{A_1, \ldots, A_n\}$ such as hand-tracking or other body-cues. Here S represents a prior distribution over objects or environment states $O = \{O_1, \ldots, O_k\}$ given by the saliency cues attached to each object . The true value of X can then be estimated using its MAP value, the value maximizing the probability

$$\overline{X} = \underset{X \in O}{\operatorname{argmax}} P(X|S, A_1, \ldots, A_n, O_1, \ldots, O_k).$$

These random variables can be encoded as a Bayesian network, as shown in Fig. 2, which makes inference much more tractable. Using the properties of such a network formulation we can then calculate the probability distribution

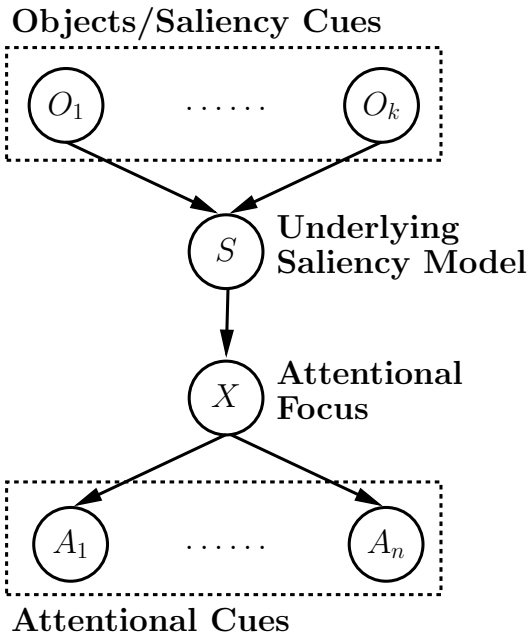


Fig. 2. **Model of Shared Attention:** This Bayesian network models the interactions of random variables that can be used to build shared attention. The set of variables $\{O_1, \ldots, O_k\}$ represent the saliency of each object given their inherent saliency cues, while X represents the current state of attention, or where the instructor is looking. The variable S denotes the underlying saliency model, a distribution over objects, and the set $\{A_1, \ldots, A_n\}$ represents attentional cues such as gaze or hand tracking.

of X given its Markov blanket, i.e.

$$\begin{aligned} P(X|S, A_{1\ldots n}, O_{1\ldots k}) &= P\left(X|\text{Blanket}(X)\right) \\ &= P\left(X|\text{Parents}(X)\right) \prod_{Z \in \text{Children}(X)} P\left(Z|\text{Parents}(Z)\right) \quad (2) \\ &= P(X|S)P(A_1|X)\cdots P(A_n|X), \quad (3) \end{aligned}$$

which scales linearly with the number of attentional cues and the complexity of the saliency model.

Given no other information, the underlying saliency model S is not directly observable. Imitation must in this case rely solely on the attentional cues, with S uniformly distributed over objects. However, as time progresses during some interaction, objects will be selected and the model updated so as to approach the true saliency distribution. This model can also be stored, on a per-instructor, per-task basis, allowing the use of previously learned information. We perform this object selection using each object's saliency cues, as detailed in section 4.

3 Using attentional cues

The process currently used to test our shared attention framework deals with close interaction between a robotic observer and a human instructor. As such we limit our attentional cues to a subset of the instructor's body, specifically gaze direction. Letting G denote the gaze vector of the instructor we can substitute into Eqn. (3), yielding

$$P(X|S,G) = P(X|S)P(G|X). \tag{4}$$

In this instance G is a 3-dimensional vector projected onto the 2-dimensional image space. We can view this as a vector with some Gaussian noise, which allows us to estimate $P(G|X)$ by computing the likelihood that attention is directed towards X given its Gaussian cloud.

The following section (3.1) details the algorithms used to track the gaze vector, while section (3.2) gives a brief overview of the uses of multiple cues.

3.1 Gaze tracking

By tracking and imitating the gaze of an instructor the observer is able to direct their own gaze to follow, in an attempt to attain shared attention. We begin this process by learning a feature-based geometric model of the head and use that to find a vector along the instructor's line of sight—a method based on work of Wu and Toyama [29]. This algorithm builds an ellipsoidal model of points, so that each point maintains the probability of local image features of the head (based on training images). Features are found in training images using a large-scale Gaussian and rotation-invariant Gabor templates at four different scales. Training is then done by building a model of the gaze direction across many example images. Relations between features found in testing images are then compared to ascertain the final directional vector.

One difficulty with this method is that it requires a tight bounding box around the head in testing and in training images. In both instances we find the instructor's head using a feature-based object detection framework developed by Viola and Jones [27]. This framework uses a learning algorithm based on AdaBoost to find efficient features and classifiers, and combines these classifiers in a cascade that can quickly discard unlikely features in test images. We use this set of algorithms because of their high detection rate and speed for detecting faces: on a standard desktop it can proceed at over 15 frames per second.

Facial identification with this framework, however, is limited in the pan and tilt angles that can still be recognized, plus or minus approximately 5–7 degrees. In order to account for this we apply our algorithms to find the face position, and continue to track the head across different movements. We apply the Meanshift algorithm [10] to track the head outside of this window of valid rotations. This is not a major limitation, as the observer also needs a signal that they've caught the attention of the instructor: seeing eye-to-eye in this case.

The observer thus begins tracking the instructor's gaze when the instructor looks at the observer, a traditional signal of attention. At this point the observer can maintain the location of the instructor's head via a tight bounding box on the instructor's face. This bounding box allows the observer to determine the instructor's gaze angle, by using the previously learned head-position model. Finally a simple Kalman filter can be used to find the most likely of these gaze-vectors.

3.2 Extension to multiple cues

Because our process involves close interaction with the instructor, we are able to limit our attentional cues to gaze tracking: whatever error is accumulated in the tracking step can be corrected using saliency cues (see section 4). There are, however, many applications in which these assumptions may not apply. Consider, for example, a mobile robot guiding a museum tour: previous work in this area has focused on touch-screen displays for interaction [8,26]. A robot of this nature might be able to use hand and arm gestures, such as pointing, to attend to specific artworks. Coupled with saliency features such as simple audio keywords (e.g. "tell me about the *Mona Lisa*") this could facilitate complex human-to-robot interaction.

Our probabilistic framework, meanwhile, provides the flexibility needed to accommodate multiple inputs. The output of each of these attentional signals should be a vector in the image space with additive Gaussian noise. The joint distribution can be found by convolving these various distributions, allowing us to easily find the MAP vector.

4 Saliency cues

In humans, shared attention via gaze following bootstraps more complex tasks, such as learning the names of objects that are the foci of attention and imitating manipulations of objects. Many sources of saliency can be used to

establish shared attention, and our framework provides a simple mechanism for combining these different sources. Our system employs a bottom-up attentional algorithm, combined with various saliency cues to build a learned model, giving an instructor-specific saliency prior over objects. This model is then combined with a top-down prior imposed by the instructor's attentional cues, as described in the previous section, to yield a context-specific estimate of the object most likely being gazed at by the instructor.

Our present system considers only one task: following the instructor's gaze to a single object. In this tracking task, the goal state s_G is achieved when observer and instructor have centered the same object in their respective visual fields. If s_G denotes a discrete-valued random variable, the distribution over objects the instructor could be looking at is $P(s_G)$. Our system begins with a single, generic model of saliency based on a biologically-inspired bottom-up attentional algorithm developed by Itti and Koch [15]. 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 subsets of the scene to be efficiently selected for higher level analysis using other saliency cues. Such an approach is mirrored in the behavior and neuronal activity of the primate visual system.

Thresholding the mask, then performing connected components on the thresholded image produces a set of discrete objects the system considers as candidates for s_G. To these candidates we can then use and convolve the various saliency cues, S_i, which with normalization will give us the distribution over objects. This distribution can then be combined with the attentional cues described in section 3 to obtain the MAP object (see Fig. 3(ghi)). Once the object has been determined, the system uses information about the object, the saliency cues, to update the instructor-specific model as described below. The final outcome of this process is a model that aims to identify, given a set of objects, a distribution over which object the instructor considers most salient to the task at hand.

As the system gathers more data on particular instructors, it builds up a context-specific model of what each instructor considers salient. One approach to this problem is to model saliency using simple features, such as color and size, which are easily extracted from images. For such an approach we can extract the size of each object and use the pixel values for color (in the YUV color space, as this provides more robustness to lighting changes). Figure 3 shows an example of this approach during both training and testing.

For each instructor, we learn a different Gaussian mixture model in YUV color space using the well known expectation maximization (EM) algorithm. In this context the EM algorithm assumes that we know the parameters of the mixture

model, and then infers the probability that each data point belongs to each Gaussian cluster. Each mixture model is trained on object pixels segmented using the bottom-up saliency method. Each training point $\mathbf{p}_i$ in this model is a vector of the form: $\mathbf{p}_i = \langle u_i, v_i, z_{i,o} \rangle$, where u_i and v_i are the UV values of pixel i, and where $z_{i,o}$ is the size of the object o (in pixels) from which pixel i was drawn.

In testing, the system uses the learned model to predict the goal states for specific instructors. The Gaussian mixture model yields a prior estimate on which object, o, the system should look at (before the instructor's attentional vectors are inferred) based on pixels in connected components. The average vector $\mathbf{p}$ over all N_x pixels in connected component x determines which Gaussian cluster connected component x is drawn from. The maximum likelihood estimate from this computation assigns a mixture component label c_o to the object. The mixture model prior for Gaussian component c_o determines the a

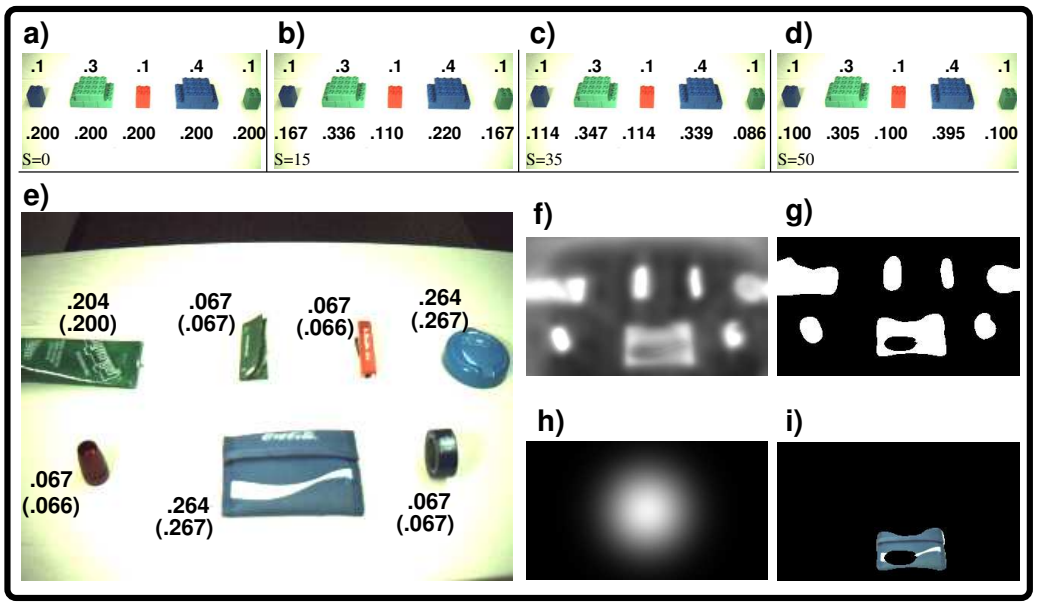


Fig. 3. **Learned saliency prior:** (a,b,c,d) The upper values give the true saliency distribution. The lower values give the current estimate for this distribution, given S samples. Progressing from (a) to (d) shows the estimate approaching the true distribution as number of samples increases. (e) After 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 [15] 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 training set (the blue wallet).

priori probability that the instructor will gaze at object o, 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:

$$c_o = \underset{c \in C}{\operatorname{argmax}} \left(\left(\frac{1}{N_x} \sum_i^{N_x} \mathbf{p}_i - \mu_c \right)^{\mathrm{T}} \Sigma_c^{-1} \left(\frac{1}{N_x} \sum_i^{N_x} \mathbf{p}_i - \mu_c \right) \right), \quad (5)$$

$$P\left(s_G = o\right) = P\left(c_o\right). \quad (6)$$

The system then combines this prior likelihood with attentional cues as described using our probabilistic model to determine an MAP estimate of where to look in 3D space.

5 Human-robot interaction

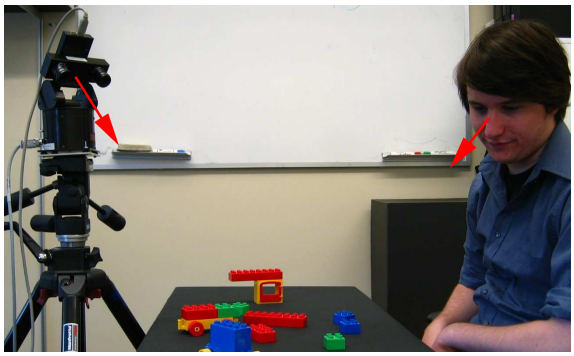
Direct interaction between a robotic observer and human instructor provides a perfect example of our approach to shared attention, and further, shows its usefulness as a tool for imitation. The robotic system used throughout this example is a Biclops active stereo vision head from Metrica, Inc. We learned a probabilistic forward model for this system by fitting a linear regression model to a set of 597 movements of the head. The model estimates encoder position error (in degrees) given an initial state and an action taken from that state. Acceleration was held to a constant 50 degrees/s^2. A cross-validation set of 896 movements established that residual error in the regression model follows a zero-mean Gaussian distribution.

5.1 Setup and gaze-tracking results

The setup for the following set of experiments involves an instructor and robotic observer (hereafter referred to as the *robot*) set at opposite ends of a table (shown in 4(a)). Initial tests within this framework focus on ascertaining the error in our gaze-tracking algorithms. The tracker was first trained using video sequences from two different instructors looking in known directions. Once completed, the tracking algorithm was tested on in- and out-of-sample instructors looking 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)).

As can be seen in Fig. 3(e), however, accurate tracking does not alleviate the problems imposed by a cluttered scene. The next set of tests deal with this

a)



b)

Gaze–tracking results:

Subj.	Succ.	Tot.	Accur.
one:	43	46	93.4%
two:	17	20	87.5%
three:	4	4	100.0%
	64	70	91.4%

Fig. 4. **Experimental setup:** Testing of the framework and algorithms espoused by this paper involves an interactive session between an instructor and a robotic observer. During the test shown in (a) the observer tracks the instructor's gaze to objects on the table and attempts to identify the most salient object. The table in (b) shows the accuracy of the gaze-tracking algorithm in distinguishing between two locations, tested with three different subjects. Only the first of these subjects is in the training set.

problem of ambiguity. The instructor and robot are still placed at opposite ends of a table, but now objects are randomly arranged on the table. The instructor has some internal saliency model (unknown to the robot) and chooses objects based on this model. Once an object has been chosen, the instructor looks towards the object, and the robot must track the instructor's gaze to the table in an attempt to determine the most salient object.

5.2 System results

Once the robot has oriented to an object in the scene, we have the robot "ask" the instructor whether they have found the correct object. This allows us to track the accuracy of our system—as time progresses the robot should correctly identify objects in fewer and fewer steps. Figure 5 plots the accuracy of our platform, where lower numbers represent more accurate object-identification. The actual values plotted are the number of object hypotheses proposed by the robot, i.e. the number of incorrect proposals plus 1. Each of the plotted tests was performed over 5 trials, with the average value being shown.

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

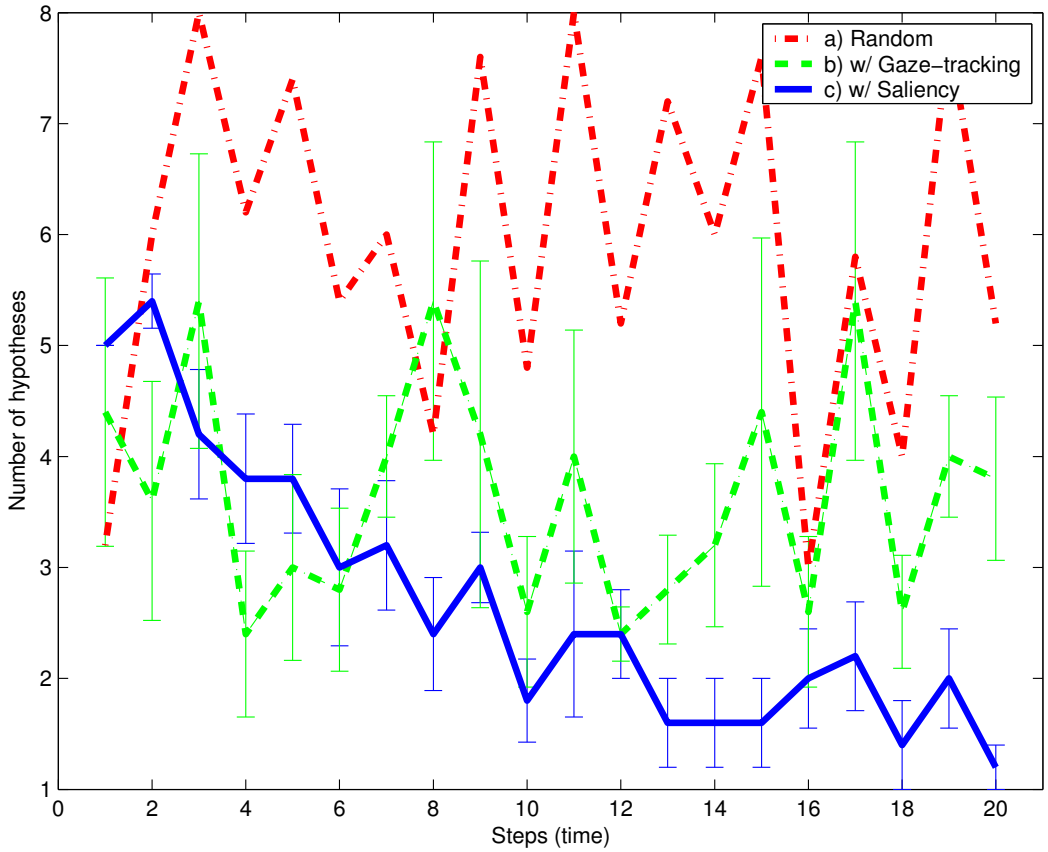


Fig. 5. **Combined results:** The above plots show the accuracy of our system at identifying 10 different objects to which the instructor is directing their attention, averaged over 5 trials. Plot a) shows the system using only random guesses as to the object, while plot b) shows the inclusion of gaze information. Plot c) combines learned saliency information with gaze tracking—beginning with a uniform prior when no model is known.

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 as time progresses the saliency approach continues to improve, with steadily declining error—while the approach using only gaze-tracking maintains the same level of accuracy.

6 Conclusion

The importance of imitation as a means to acquire knowledge and skills has been recognized by a growing number of researchers in the robotics community. Two such researchers include Breazeal and Scassellati who in [5] lay out an approach to robotic imitation and outline the requirements for an imitative system. They 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 [6] (and more recently with Leonardo [4]). 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. Their work, however, lacks some features for performing complex imitation tasks. The approach espoused by their work does not appear to employ a single unifying framework or mathematical formalism for imitation. Kismet's attention seems deterministically driven, with fixed responses and expressions, limiting its applicability in less controlled environments. This lack of a unifying framework makes their system difficult to compare and contrast with results from the cognitive literature.

Earlier research on head and gaze imitation has been performed by Demiris et al [11]. This work, however, is limited to gaze imitation with no capacity for shared attention; the system merely mimics the instructor's head position and makes no attempt to follow their gaze. The work of Nagai et al in [21] 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. This, however, presents a limited model of shared attention, and making it difficult to include further information: hand gestures, audio cues, etc. Our system, further, includes a Bayesian network model of shared attention which allows for the inclusion of various sources of data.

This paper presents a Bayesian framework for imitation learning, and shows how shared attention fits into the framework. The framework builds on Meltzoff and Moore's AIM hypothesis for human imitative acts. We anticipate extending our saliency learning and gaze tracking system to the HOAP-2 humanoid platform in the near future. Our algorithmic framework is hardware-agnostic, except for the forward model; instructor head pose estimation and the prior model will not change under this platform. Once we learn the forward dynamics of the humanoid's head, gaze following and saliency model learning will employ the same codebase as the Biclops head. This extension will in turn enable more complex imitative tasks to be learned. We believe that this method can be put to greater use in task-specific environments. Using the current set of LEGO™ objects, this could be a task such as "build a fire truck".

Such a task would involve different sizes and shapes of building-blocks, with a predominance of red blocks, allowing an easy-to-understand model of saliency for the Biclops to learn. We also anticipate expanding our saliency learning system to accommodate more attentional cues (such as auditory information and pointing) and richer saliency models.

References

[1] A. Billard and M. J. Mataric. A biologically inspired robotic model for learning by imitation. In C. Sierra, M. Gini, and J. S. Rosenschein, editors, *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 373–380, Barcelona, Catalonia, Spain, 2000. ACM Press.

[2] S. J. Blakemore, S. J. Goodbody, and D. M. Wolpert. Predicting the consequences of our own actions: the role of sensorimotor context estimation. *J. Neurosci.*, 18(18):7511–7518, 1998.

[3] C. Breazeal. Imitation as social exchange between humans and robots. In *Proc. AISB99*, pages 96–104, 1999.

[4] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning from and about others: towards using imitation to bootstrap the social understanding of others by robots (to appear). *Artificial Life (special issue)*, 2004.

[5] C. Breazeal and B. Scassellati. Challenges in building robots that imitate people. In K. Dautenhahn and C. Nehaniv, editors, *Imitation in Animals and Artifacts*. MIT Press, 2001.

[6] C. Breazeal and J. Velasquez. Toward teaching a robot "infant" using emotive communication acts. In *Proc. 1998 Simulation of Adaptive Behavior, Workshop on Socially Situated Intelligence*, pages 25–40, 1998.

[7] R. Brooks and A. Meltzoff. The importance of eyes: How infants interpret adult looking behavior. *Dev. Psych.*, 38:958–966, 2002.

[8] J. M. Buhmann, W. Burgard, A. B. Cremers, D. Fox, T. Hofmann, F. E. Schneider, J. Strikos, and S. Thrun. The mobile robot RHINO. *AI Magazine*, 16(2):31–38, 1995.

[9] R. W. Byrne and A. E. Russon. Learning by imitation: a hierarchical approach. *Behavioral and Brain Sciences*, 2003.

[10] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 142–151.

[11] J. Demiris, S. Rougeaux, G. Hayes, L. Berthouze, and Y. Kuniyoshi. Deferred imitation of human head movements by an active stereo vision head. In *Proc. of the 6th IEEE International Workshop on Robot Human Communication*, 1997.

[12] I. Fasel, G. O. Deak, J. Triesch, and J. R. Movellan. Combining embodied models and empirical research for understanding the development of shared attention. In *Proc. ICDL 2*, 2002.

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

[14] M. Haruno, D. Wolpert, and M. Kawato. MOSAIC model for sensorimotor learning and control. *Neural Computation*, 13:2201–2222, 2000.

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259, 1998.

[16] M. I. Jordan and D. E. Rumelhart. Forward models: supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

[17] J. R. Krebs and A. Kacelnik. Decision making. In J. R. Krebs and N. B. Davies, editors, *Behavioural Ecology (3rd ed.)*, pages 105–137. Blackwell Scientific Publishers, 1991.

[18] M. Lungarella and G. Metta. Beyond gazing, pointing, and reaching: a survey of developmental robotics. In *EPIROB '03*, pages 81–89, 2003.

[19] A. N. Meltzoff and M. K. Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78, 1977.

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

[21] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. Emergence of joint attention based on visual attention and self learning. In *Proc. of 2nd International Symposium on Adaptive Motion of Animals and Machines*, 2003.

[22] B. Price. *Accelerating Reinforcement Learning with Imitation*. PhD thesis, University of British Columbia, 2003.

[23] R. P. N. Rao and A. N. Meltzoff. Imitation learning in infants and robots: Towards probabilistic computational models. In *Proc. AISB*, 2003.

[24] R. P. N. Rao, A. P. Shon, and A. N. Meltzoff. A Bayesian model of imitation in infants and robots. In *Imitation and Social Learning in Robots, Humans, and Animals*. Cambridge University Press, 2004 (to appear).

[25] B. Scassellati. 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, 1999.

[26] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. MINERVA: A second-generation museum tour-guide robot. In *Proceedings: IEEE International Conference on Robotics and Automation (ICRA '99)*, Detroit, Michigan, May 1999.

[27] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.

[28] E. Visalberghy and D. Fragaszy. Do monkeys ape? In *Language and intelligence in monkeys and apes: comparative developmental perspectives*, pages 247–273. 1990.

[29] Y. Wu, K. Toyama, and T. Huang. Wide-range, person- and illumination-insensitive head orientation estimation. In *AFGR00*, pages 183–188, 2000.