Probabilistic Gaze Imitation and Saliency Learning in a Robotic Head

Aaron P. Shon, David B. Grimes, Chris L. Baker, Matthew W. Hoffman, Shengli Zhou, and Rajesh P.N. Rao
{aaron,grimes,clbaker,mhoffman,shengliz,rao}@cs.washington.edu
CSE Department, Box 352350 University of Washington Seattle WA 98195 USA

***Abstract*— Imitation is a powerful mechanism for transferring knowledge from an instructor to a naïve observer, one that is deeply contingent on a state of shared attention between these two agents. In this paper we present Bayesian algorithms that implement the core of an imitation learning framework. We use gaze imitation, coupled with task-dependent saliency learning, to build a state of shared attention between the instructor and observer. We demonstrate the performance of our algorithms in a gaze following and saliency learning task implemented on an active vision robotic head. Our results suggest that the ability to follow gaze and learn instructor- and task-specific saliency models could play a crucial role in building systems capable of complex forms of human-robot interaction.**

I. Imitation learning and shared attention

Imitation is a powerful mechanism for transferring knowledge from a skilled agent (the *instructor*) to an unskilled agent (or *observer*) using direct manipulation of the environment. Several researchers have investigated imitative behavior in apes [1], [2], in children (including infants only 42 minutes old) [3], [4], and in an increasingly diverse selection of machines [5], [6]. The attraction of imitation for robotics is obvious: imitative robots offer drastically reduced programming costs compared to robots requiring programming by an expert. Imitative robots also offer testbeds for cognitive researchers to test computational theories, and provide modifiable agents for contingent interaction with humans in psychological experiments.

Successful imitation requires that instructor and observer simultaneously attend to the same object or environmental state. Such simultaneous attention is often referred to as "shared attention" in the psychological literature. Previous work by Scassellati focused on tracking the gaze of a human instructor, and on mimicking the motion of the instructor's head in either a vertical or a horizontal direction. Separately, Triesch and colleagues have used robotic platforms to study shared attention in infants [7].

Although robotic systems [8], [9] have demonstrated impressive mimicry results, richly contingent human-robot interaction comparable to infant 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.

In this paper, we describe a robotic system that uses probabilistic algorithms to follow the gaze of a human and identify salient 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. Bayesian imitation learning approaches have been proposed to accelerate reinforcement learning [10]; however, that framework chiefly addresses the problem of learning a forward model of the environment [11] via imitation (see Section IV), and its correspondence with cognitive findings in humans is unclear. Other frameworks have been proposed for imitation learning in machines [9], [12], [13], but most of these are not designed around a coherent probabilistic formalism.

The robotic system described in this paper tracks a human instructor's gaze to an object, then learns a simple instructor-specific, 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 framework; and iii) the system learns simple, context-dependent probabilistic models for saliency. Our preliminary results show how shared attention could be developed between humans and robots.

Section II describes the underlying software architecture used for our research. Sections III and IV respectively discuss our system's modality-independent representation and our Bayesian algorithms for motor planning. Section V describes how our system computes object saliency. Section VI concludes with gaze tracking results from our system and a simple example of learning a saliency model.

II. System architecture

Our robotic system is a Biclops active stereo vision head from Metrica, Inc. (see Fig. 2(b)). The head's two cameras provide 30 fps color video at 320×240 resolution. The motor component consists of two servos and optical position encoders for pan and tilt control.

Our software architecture seeks to abstract away details of low level hardware, making it easy to develop modular services which perform information gathering or processing functions. Services can subscribe to other services, allowing a chain of "filters," each of which can run on a different machine. A modular, distributed architecture is a natural fit with real-time robotic systems which operate on the principle of selecting actions based on Bayesian inference. Services in our architecture generally communicate by passing serialized

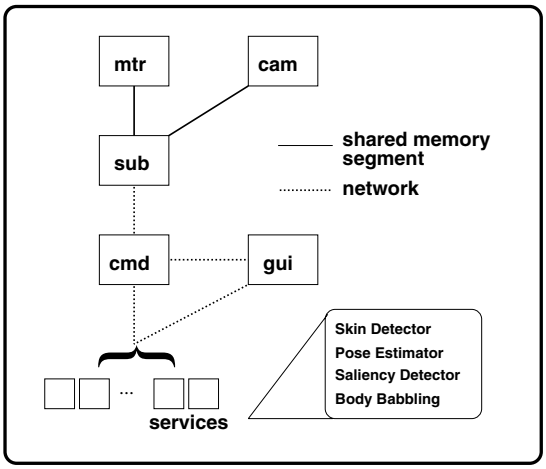

Fig. 1. **System architecture:** The `mtr` and `cam` processes connect directly to the Biclops hardware, and are responsible for transmitting motor states and images respectively. These objects are encapsulated and sent to `sub`, which acts as a gateway between the system and the hardware. The objects are then sent along a network connection to `cmd`, which is able to redistribute the objects to the waiting services, and to allow state display through the gui.

representations of probability distributions. The distributed nature of the architecture allows for a responsive, real-time interaction between the human instructor and the robotic head. In our experiments, computation is distributed among three commodity Pentium workstations.

The software architecture is divided into two layers, an initial layer which directly communicates with the hardware, and a layer of services which act on this data. In between these two steps is the `sub` process, which acts as a gateway between the two halves. The initial layer is split into two main processes, `mtr` and `cam`, which obtain and serialize output from the position encoders and cameras respectively. The `mtr` process also accepts motor commands and passes them along to the servo control hardware.

All data is serialized before being transmitted to `sub`, and this provides a means for network transparent data transfer between the later components of our architecture. Once `cam` and `mtr` have serialized their data, the serialized objects are transmitted to `sub` via a shared memory segment. `sub` then transmits the objects over the network to `cmd`, a process which distributes each individual object to the running services. `sub` also communicates with the `gui` process, which displays the images and provides a visual control mechanism.

III. State Estimation and Modality-Independent Representation

Meltzoff and Moore's Active Intermodal Mapping (AIM) hypothesis [4] 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. AIM informs the structure of our robotic system. Fig. 2 juxtaposes the elements of AIM and our system.

At the beginning of each trial, a histogram-based color detection system [14] first finds regions likely to contain human skin. Each connected component is annotated with a bounding box. The maximum a posteriori (MAP) bounding box is selected given a prior over bounding box size and aspect ratio. Selecting the MAP hypothesis is a common approximation, where predictions are made on a single most probable hypothesis. Kalman filtering stabilizes the box location over successive frames. After finding a stable bounding box containing the instructor's face $\mathbf{b}_I^*$, a probabilistic algorithm [15] computes a distribution over instructor's head pose $P(\mathbf{h}_I|\mathbf{b}_I^*)$. This algorithm finds edge-density and texture features within the bounding box and compares these features to a previously trained geometric model of the head.

Note that the system does not infer gaze from the instructor's eyes, a development that only occurs past age 9 months in infants [16], well past the onset of imitative behavior (which in some cases [3] is present from birth).

Finally for each head pose we compute a gaze vector and project the head pose distribution $P(\mathbf{h}_I|\mathbf{b}_I^*)$ into a simplified 3D model of the world using estimated intrinsic and extrinsic camera parameters. The resulting distribution $P(s_G|\mathbf{b}_I^*)$ forms the inter-modal representation of the common goal of Biclops and instructor.

IV. Probabilistic Forward, Inverse, and Policy Models

Many robotics tasks model the environment, whether using a static map of an area or running a dynamical simulator of the world over time. Forward and inverse models [11] provide

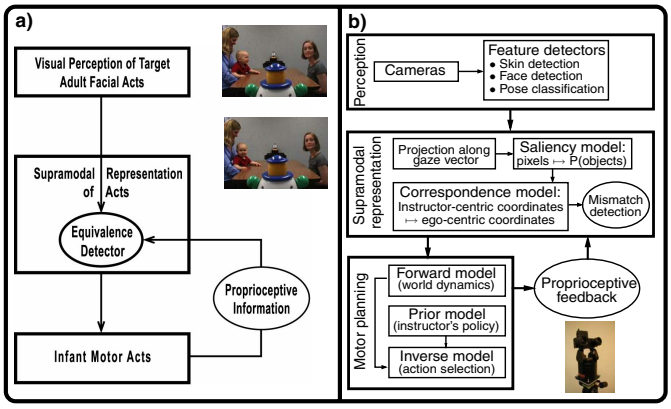

Fig. 2. **AIM hypothesis for infant imitation:** (a) The AIM hypothesis of facial imitation by Meltzoff and Moore [4] argues that infants match observations of adults with their own proprioception using a modality-independent representation of state. Mismatch detection between infant and adult states 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 [16]). (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.

a framework for using 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. Wolpert and colleagues have modeled paired forward and inverse models for motor control and imitation, and investigated their neurological implementations [17], [18].

Reinforcement learning systems typically acquire a third type of model, which we call a policy model. Policy models compute distributions over actions that an agent should take to reach a goal, given the current state of the environment and the agent itself. Let s_t be the combined state of the environment and an agent in the environment at time t, let s_G be a goal state the agent wishes to achieve (perhaps representing a state of high-valued reward in a reinforcement learning framework), and let a_t be an action taken at time t. Assuming a first-order Markovian environment, a probabilistic forward model can be represented as $P(s_{t+1}|a_t, s_t, s_G) \equiv P(s_{t+1}|a_t, s_t)$, and the corresponding inverse model can be represented as $P(a_t|s_t, s_{t+1}, s_G)$. Similarly, a policy model can be denoted as $P(a_t|s_t, s_G)$.

Learning an inverse model is the desired outcome for a learning agent that wishes to imitate, 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. 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 α) [19], [20]:

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

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" [21], which seems to be used in at least some cases by the brain. Our present system uses only MAP estimates to select actions.

We learned a probabilistic forward model for the Biclops by fitting a linear regression model to encoder position error (in degrees) given an initial state, s_t, and an action taken from that state, a_t; acceleration was held to a constant 50 degrees/s^2. Fig. 3(a) shows error values, $|s_G - s_{t+1}|$, from 597 training movements; Figs. 3(b,c) show that remaining error is marginally Gaussian. Fig. 3(d) shows cross-validation of the model using a testing set of 896 movements.

Learning a policy model $P(a_t|s_t, s_G)$ requires inferring actions a_t based on the instructor's state transitions. This inference from state transitions to actions in turn requires knowing the "action inference" distribution $P(a_t|s_t, s_{t+1})$. A full-fledged Bayesian approach to learning policy models would propagate the uncertainty in this estimate through the policy model.

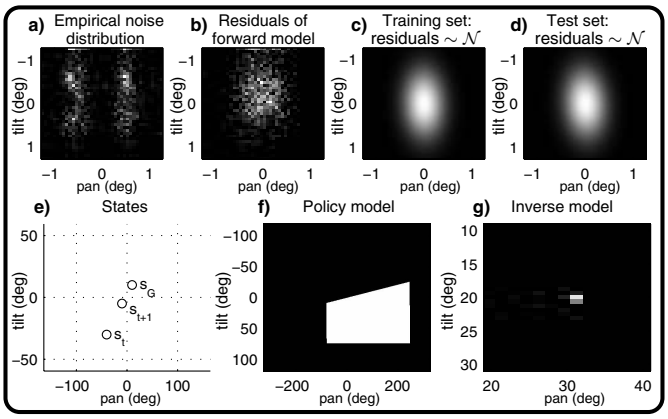

Fig. 3. **Probabilistic forward model:** (a) Discrete empirical distribution of deviation (in degrees) between intended motor states and observed motor states in the Biclops stereo head. The distribution was estimated using a training dataset of 597 example movements. If motors in the head were completely accurate, the entire distribution would display a spike at (0,0). (b) Distribution of residual error after model learning. A linear model was learned in a maximum likelihood fashion from the training data. The residual error distribution is marginally Gaussian, an important assumption of the model. (c) Gaussian approximation of the discretized distribution shown in (b). (d) Gaussian approximation of errors on a cross-validation set of 896 testing movements resembles that of the training set. (e-g) Example of an inverse model computation. (e) States used in this example were: $s_t = (-40, -30)$, $s_{t+1} = (-10, -10)$, and $s_G = (10, 10)$. (f) Policy model $P(a_t|s_t, s_G)$. The grid shows depicts the policy that only actions yielding next states closer to the goal than the current state are allowed, i.e. given non-zero probability by the policy model. (g) Inverse model $P(a_t|s_t, s_{t+1}, s_G)$. The distribution shows the likelihood of each action to move the Biclops head toward s_{t+1}. This distribution is sharply peaked around $a_t = (30.36, 19.63)$.

The present system does not learn a policy model. 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. 3(f) shows the prior model $P(a_t|s_t, s_G)$ conditioned on $s_t = (-40, -30)$ and $s_G = (10, 10)$ (as depicted in Fig. 3(e)).

Finally, Fig. 3(g) shows the inverse model $P(a_t|s_t, s_{t+1}, s_G)$ conditioned on $s_t = (-40, -30)$, $s_{t+1} = (-10, -10)$, and $s_G = (10, 10)$. The system then selects the MAP action to move the Biclops head to s_{t+1}, thus combining the information from the prior and forward models.

V. Modeling Saliency

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. Our system employs 3 image-based sources: i) a bottom-up attentional algorithm; ii) a top-down prior imposed by the instructor's gaze vector, computed as described in the previous section; and iii) a learned model that gives an instructor-specific saliency prior over objects. These 3 saliency cues combine to yield a context-specific estimate of the object most likely being gazed at by the instructor. In the future, we envision combining auditory cues (e.g., "look at the large

red object") with the other 3 sources to increase attentional fidelity.

Our present results consider only one task: gaze following to a single salient object. In tracking the instructor's gaze to an object, 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 $P(s_G)$. This distribution intuitively corresponds to saliency: objects the instructor considers relevant to a task are more likely to be fixated on the instructor. Our system begins with a single, generic model of saliency based on a biologically-inspired bottom-up attentional algorithm [22]. This algorithm returns a saliency "mask" (see Fig. 4(f)) where the grayscale intensity of a pixel is proportional to saliency as computed from feature detectors for intensity gradients, color, and edge orientation.

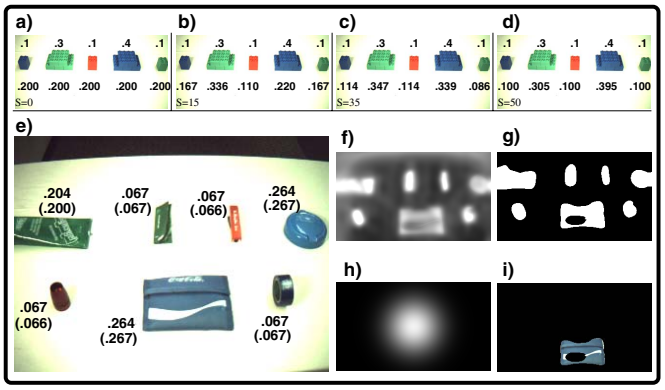

Fig. 4. **Learned saliency prior:** (a,b,c,d) The upper values give the true saliency distribution. The lower values give the current estimate for this distribution, given S samples. Progressing from (a) to (d) shows the estimate approaching the true distribution as number of samples increases. (e) After 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 [22] 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).

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. During training, the system uses the instructor's estimated gaze vector to disambiguate between candidate objects. Once the object gazed at by the instructor is determined, the system uses information about the object to learn an instructor-specific saliency 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. Our initial approach is to model instructor-specific saliency with simple features such as color and size which are easily extracted from images. However this model is easily extensible, and in the future we intend to utilize higher order features such as shape, and object categories. 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 $\mathbf{p}_i$ to the 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. Together, these distributions model the saliency preferences of the instructor.

The EM algorithm for learning Gaussian mixture models provides a simple, robust method for clustering data, and for this reason was chosen as the basis for our saliency model. Our tests demonstrate (Fig. 4) that this method is able to quickly approach the true distribution. Fig. 5(h) further shows the improvement of this approach over a uniform saliency model.

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 gaze vector is 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 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:

$$c_o = \underset{c \in C}{\operatorname{argmax}} \left(\left(\frac{1}{N_x} \sum_i^{N_x} \mathbf{p}_i - \mu_c \right)^{\mathrm{T}} \Sigma_c^{-1} \left(\frac{1}{N_x} \sum_i^{N_x} \mathbf{p}_i - \mu_c \right) \right) \quad (2)$$

$$P(s_G = o) = P(c_o) \quad (3)$$

The system combines this prior likelihood with likelihoods given by the instructor's gaze vector to determine an MAP estimate of where to look in 3D space.

VI. Results: Gaze Tracking and Saliency Model Learning

During training, we placed 5 distinct objects on the table. On each trial, an instructor looked at the 5 objects in the scene according to some underlying (hidden) probability distribution. The Biclops then tracked the instructor's gaze down to the table, assumed the salient object was located at the center of the resulting image, and updated its estimate of the instructor's saliency model accordingly. Figs. 4(a,b,c,d) show saliency model learning at four different points in the training process. Fig. 4(a) plots the model's saliency estimate (lower row of

text) as a distribution over objects before training begins (with $S = 0$ training examples from the instructor). The true distribution the instructor used to select objects is shown in the upper row of text. In Fig. 4(b,c,d), as more training samples are collected from the instructor ($S = 15$, $S = 35$, and $S = 50$), the estimated saliency distribution becomes closer to the true distribution. The instructor shown here prefers large green and large blue objects. Fig. 4(e,f) respectively show the testing performance and the grayscale saliency map given by the bottom-up algorithm. The testing objects are distinct from the training objects, but share similar surface colors and object sizes. Note that the saliency distribution estimated by the model on the testing objects intuitively matches the instructor preferences shown during training—the model assigns large blue objects much higher probabilities of being salient compared to other object types. Fig. 4(g) shows the thresholded saliency map, while Fig. 4(h) shows the intersection of the instructor's gaze vector and the table with additive Gaussian noise. These are combined with the likelihoods from Fig. 4(e) to give an MAP estimate of the most salient object, shown in Fig. 4(i).

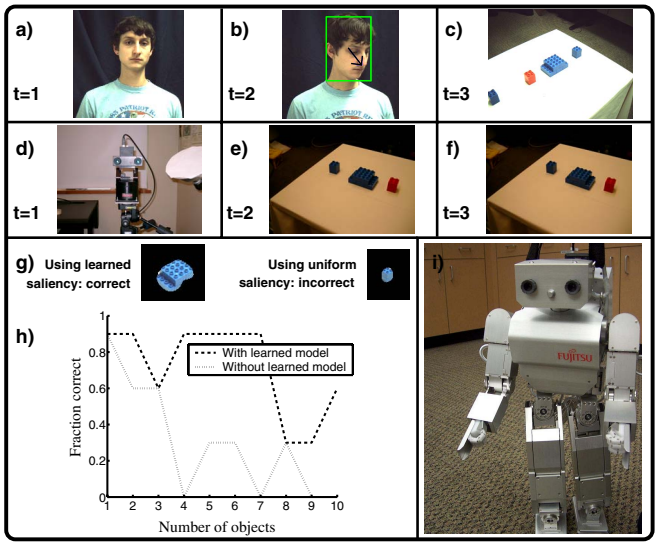

Fig. 5. **Gaze following performance:** (a-f) Testing process. Top row shows the Biclops' view, bottom row shows the instructor's view. From left to right, the Biclops finds the instructor's face and annotates an estimated gaze vector. Next the Biclops looks at the table. In (c), the Biclops correctly infers the object attended to by the instructor by combining the estimated gaze vector and estimated object saliency. (g) Using learned saliency prior enables disambiguation of relevant objects from a cluster of closely situated objects. A uniform saliency prior fails, while the learned saliency model shown in Fig. 4(a-d) identifies the correct object. (h) In this line graph, the dark line represents the performance of the Biclops with the learned saliency model, and the lighter line represents performance with no saliency. At each point the accuracy of the Biclops averaged over 3 trials is plotted against the number of objects in the scene. (i) HOAP-2 humanoid robot. Our future efforts will focus on sensorimotor learning and imitation learning on this platform.

Fig. 5(a-f) show the testing process. The top row shows the process from the Biclops' viewpoint, while the bottom row shows the instructor's viewpoint. From left to right, the Biclops first infers the instructor's gaze vector, follows the gaze vector to a cluster of objects, and centers on the most salient object. The combined pose estimation algorithm and action selection yield 90% accuracy on matching the instructor's gaze vector using out-of-sample data. Fig. 5(g) demonstrates the value of an instructor-specific saliency model: when the instructor's gaze tracks to a cluster of objects that the bottom-up algorithm regards as salient (i.e., when instructor gaze contains ambiguity), a learned saliency prior enables the system to select the instructor's object of interest more often than using a uniform prior over object saliency.

We performed additional tests to ascertain the accuracy of the saliency subsystem in an increasingly cluttered scene. Given a previously learned instructor-specific model of saliency, random objects were added to the scene and the instructor was asked to look at the scene and pick out an object. The Biclops then viewed the same scene, and picked out its estimate for the most salient object. The chosen object was then matched against the instructor's choice for most salient object. We repeated this process until the scene contained ten objects. Finally, we repeated the entire test three times. The line graph in Fig. 5(h) contrasts performance of the learned saliency model with the system using no learned model, that is, where each object type is uniformly likely. As the number of potentially salient objects in the instructor's gaze vector increases, the instructor's gaze vector becomes increasingly ambiguous as a marker of which object the instructor considers salient. The learned saliency model continues to robustly identify the object at which the instructor is gazing over increasing number of objects, while performance using the uniform prior quickly degrades.

VII. 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 [23] 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 [24] (and more recently with Leonardo [25]). 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 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 [26]. 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 [27] 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 an implicit Bayesian network (a trivial network in this paper), which allows us to include various sources of data. (See [28] for further information.)

This paper presents a Bayesian framework for imitation learning, and shows how gaze following to salient objects fits into the framework. The framework builds on Meltzoff and Moore's AIM hypothesis for human imitative acts. Preliminary results from an active vision stereo head demonstrate the ability of our system to learn simple saliency preferences, and to track instructor gaze to salient objects. We anticipate extending our saliency learning and gaze tracking system to the HOAP-2 humanoid platform (Fig. 5(h)) 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.

ACKNOWLEDGMENT

We thank Andy Meltzoff for providing Fig. 2(a). Thanks also to RWTH-Aachen for spearheading development of LTI-Lib, used as the basis for many of our algorithmic implementations (see `http://ltilib.sf.net/doc/homepage` for details). This work was supported by NSF grant no. 0413335 and ONR grant no. N14-03-1-0457.

REFERENCES

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

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

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

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

[5] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 142–166, 2002.

[6] M. Lungarella and G. Metta, "Beyond gazing, pointing, and reaching: a survey of developmental robotics," in *EPIROB '03*, 2003, pp. 81–89.

[7] 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.

[8] J. Demiris and G. Hayes, "A robot controller using learning by imitation," in *Proc. ISIRS*, 1994.

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

[10] B. Price, "Accelerating reinforcement learning with imitation," Ph.D. dissertation, University of British Columbia, 2003.

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

[12] C. Breazeal, "Imitation as social exchange between humans and robots," in *Proc. AISB99*, 1999, pp. 96–104.

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

[14] R. T. Collins and Y. Liu, "On-line selection of discriminative tracking features," in *Proc. ICCV 9*, 2003, pp. 346–352.

[15] Y. Wu, K. Toyama, and T. Huang, "Wide-range, person- and illumination-insensitive head orientation estimation," in *AFGR00*, 2000, pp. 183–188.

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

[17] 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.*, vol. 18, no. 18, pp. 7511–7518, 1998.

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

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

[20] 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).

[21] J. R. Krebs and A. Kacelnik, "Decision making," in *Behavioural Ecology (3rd ed.)*, J. R. Krebs and N. B. Davies, Eds. Blackwell Scientific Publishers, 1991, pp. 105–137.

[22] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

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

[24] 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*, 1998, pp. 25–40.

[25] 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.

[26] 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.

[27] 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.

[28] M. W. Hoffman, A. P. Shon, D. B. Grimes, C. L. Baker, and R. P. N. Rao, "A Bayesian active vision architecture for shared attention," University of Washington, Tech. Rep. 2005-01-01, 2005.