University of Washington CSE TR 2004-07-02
Submitted to NIPS 2004

Probabilistic Gaze Imitation and Saliency Learning in a Robotic Head

Aaron P. Shon, David B. Grimes, Chris L. Baker,
Matthew W. Hoffman, Shengli Zhou, 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 naive observer. We first present Bayesian algorithms, based on Meltzoff and Moore's AIM model for imitation in infants, that implement the core of an imitation learning framework. Next, we present Bayesian algorithms for learning which objects an instructor considers salient in a task. Finally, 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.

1 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 [21, 5], in children (including infants only 42 minutes old) [15, 16], and in an increasingly diverse selection of machines [9, 14]. 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, notably by Scassellati on the Cog platform [20], has concentrated on deterministic algorithms for shared attention between humans and robots. Scassellati's work concentrated 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, Movellan and colleagues have used robotic platforms to study shared attention in infants [8].

Although robotic platforms [7, 20] 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 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 to 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 [17]; however, that framework chiefly addresses the problem of learning a forward model of the environment [12] via imitation (see Section 3), and its correspondence with cognitive findings in humans is unclear. Other frameworks have been proposed for imitation learning in machines [3, 20, 1], but most of these are not designed around a coherent probabilistic formalism.

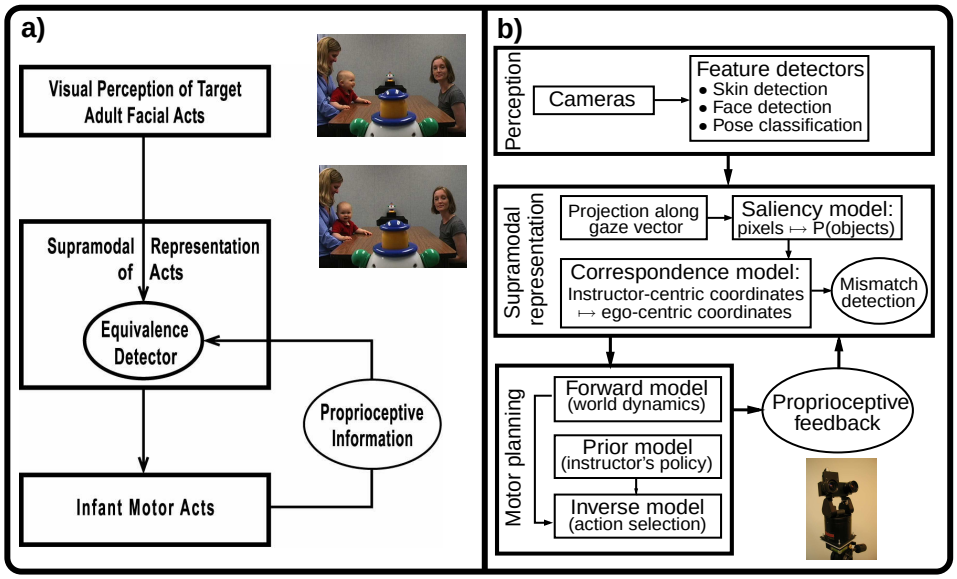

Figure 1: **AIM hypothesis for infant imitation:** (a) The AIM hypothesis of facial imitation by Meltzoff and Moore [16] argues that infants match observations of adults with their own proprioceptions using a modality-independent representation of state. Mismatch detection between infant and adult states 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 [4]). (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.

gions likely to contain human skin. Each connected component is annotated with a bounding box. The MAP bounding box is selected given a prior over bounding box size and aspect ratio. Kalman filtering stabilizes the box location over successive frames. After finding a stable bounding box containing the instructor's face, a probabilistic algorithm [22] extracts a maximum-likelihood estimate for the instructor's head pose $\mathbf{h}_I$. The system does not infer gaze from the instructor's eyes, a development that only occurs past age 9 months in infants [4], well past the onset of imitative behavior (which in some cases [15] is present from birth). Finally from $\mathbf{h}_I$ we compute a 3D gaze vector $\mathbf{v}_I$ for the instructor. We project the gaze vector $\mathbf{v}_I$ into a simplified 3D model of the world using estimated instrinsic camera parameters. This space forms the inter-modal representation of the common goal of Biclops and instructor.

3 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 [12] provide 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 [2, 10].

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 k) [19, 18]:

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

Actions can be selected in one of two ways given such an inverse model. The observer can select the action with maximum likelihood, or the observer can sample from $P(a_t|s_t, s_{t+1}, s_G)$, a strategy known as "probability matching" [13], which seems to be used in at least some cases by the brain. Our present system uses only maximum likelihood 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 and an action taken from that state; acceleration was held to a constant 50 degrees/s^2. Fig. 2(a) shows error values from 597 training movements; Fig. 2(b,c) show that remaining error is marginally Gaussian. Fig. 2(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 $P(a_t|s^\iota_t, s^\iota_{t+1})$, where s^ι_t refers to a subset of instructor's motor state at time t. A full-fledged Bayesian approach to learning policy models would propagate the uncertainty in this estimate through the policy model.

The present system does not learn a policy model. Instead, the system 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 ML 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. 2(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. 2(e)). Finally, Fig. 2(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 selects the maximum likelihood action to move the Biclops head to s_{t+1}. The prior model and forward model combine to yield an action estimate close to the displacement between s_t and s_{t+1}.

4 Modeling saliency

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

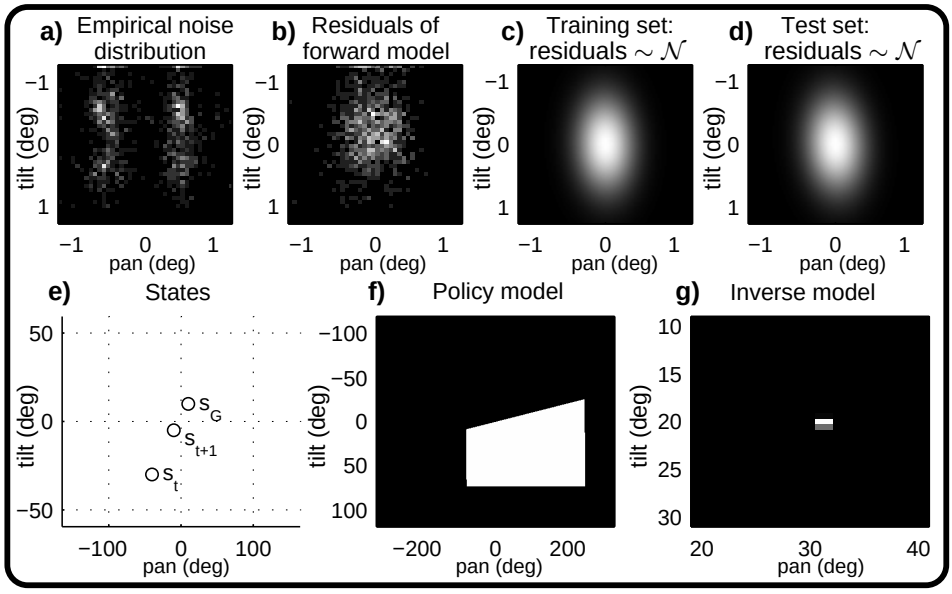

Figure 2: **Probabilistic forward model:** (a) Discrete empirical distribution of deviation (in degrees) between intended motor states and observed motor states in the Biclops stereo head. The distribution was estimated using a training dataset of 597 example movements. If motors in the head were completely 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 $a_t = (30.36, 19.63)$.

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 [11]. 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.

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.

Implicitly, the bottom-up algorithm integrates out any instructor-specific and task-specific

saliency information:

$$P(s_G) = \sum_{I \in \text{instructors}} \sum_{T \in \text{tasks}} P(s_G|I,T)\, P(I,T) \tag{2}$$

As the system gathers more data on particular instructors, it builds up a context-specific model of what each instructor considers salient. For each instructor, we learn a different Gaussian mixture model in YUV color space using the expectation maximization (EM) algorithm. 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.

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 likelihood 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:

$$c_o = \underset{c \in C}{\operatorname{argmax}} \left(\left(\frac{1}{N_x} \sum_i^{N_x} \mathbf{p}_i - \mu_c \right)^{\mathrm{T}} \Sigma_c^{-1} \left(\frac{1}{N_x} \sum_i^{N_x} \mathbf{p}_i - \mu_c \right) \right) \tag{3}$$

$$P(s_G = o) = P(c_o) \tag{4}$$

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.

5 Results: gaze tracking and saliency model learning

Fig. 3(a,b,c,d) show saliency model learning at four different points in the training process. Fig. 3(a) plots the model's saliency estimate (upper row of text) as a distribution over objects before training begins (with $S = 0$ gaze examples from the instructor). The true distribution the instructor (Fig. 3(e)) used to select objects is shown in the lower row of text. In Fig. 3(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. 3(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(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. 4(g) demonstrates the value of an instructor-specific saliency model: when the instructor's gaze tracks to a cluster of objects the bottom-up algorithm regards as salient (that is, 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. The line graph in Fig. 4(g) contrasts performance of the combined pose detector, inverse model, and instructor-specific saliency model (dashed line) with the pose

detector, inverse model, and a uniform prior on object saliency (solid line). 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[1].

6 Conclusion

This paper presented a Bayesian framework for imitation learning, and showed 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 demonstrated 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. 4(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 under our framework. We also anticipate expanding our saliency learning system to accommodate more attentional cues (such as auditory information) and richer saliency models.

Acknowledgements: We thank Andy Meltzoff for generously providing Fig. 1(a). This work was supported by grants from NSF and ONR.

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

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

[6] R. T. Collins and Y. Liu. On-line selection of discriminative tracking features. In *Proc. ICCV 9*, pages 346–352, 2003.

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

[8] 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.

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

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

[11] 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.

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

[13] 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.

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

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

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

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

[1] Please see `http://neural.cs.washington.edu/` for movies of the Biclops imitating.

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

[19] 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).

[20] 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.

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

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

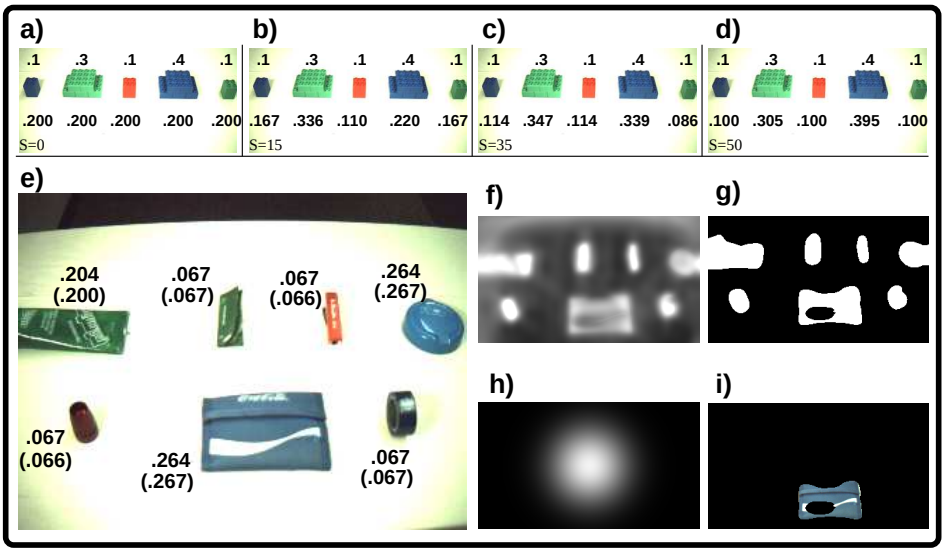

Figure 3: **Learned saliency prior:** (a,b,c,d) The upper values give the true saliency distribution. The lower values give the current estimate for this distribution, given S samples. Progressing from (a) to (d) shows the estimate approaching the true distribution as number of samples increases. (e) 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 [11] 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 an MAP estimate for the most salient object in the training set (the blue wallet).

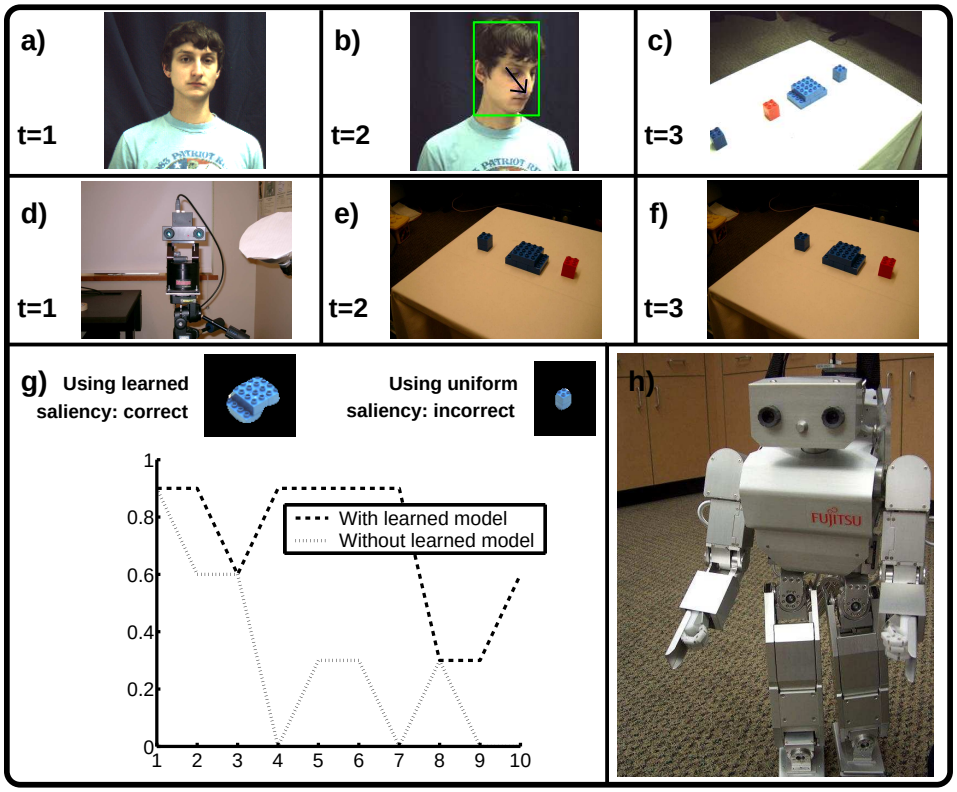

Figure 4: **Gaze following performance:** (a-f) Testing process. Top row shows the Biclops' view, bottom row shows the instructor's view. From left to right, the Biclops finds the instructor's face and annotates an estimated gaze vector. Next the Biclops looks at the table. In (c), the Biclops correctly 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. 3(a-d) identifies the correct object. As number of objects in the gaze vector increases, the learned saliency model (dashed line) outperforms a uniform saliency prior (solid line). The plot shows an average over 3 trials for both the learned saliency and uniform saliency cases. (h) HOAP-2 humanoid robot. Our future efforts will focus on sensorimotor learning and imitation learning on this platform.