Building a Model of Infant Social Interaction

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
josh@cogsci.ucsd.edu

Gedeon O. Deák
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
deak@cogsci.ucsd.edu

Hector Jasso
hmjasso@gmail.com

Jochen Triesch
Frankfurt Institute for Advanced Studies
Johann Wolfgang Goethe University
60438 Frankfurt am Main, Germany
triesch@fias.uni-frankfurt.de

Abstract

Naturalistic observations of infant/caregiver social attention have yielded rich information about human social development. However, observational data are expensive, laborious, and reliant on fallible human coders. We model interactions between caregivers and infants using a three dimensional simulation environment in order to gain greater insight into the development of infant attention sharing, specifically gaze following. Most models of infant cognition have been only abstractly linked to the detail of a real life environment and to the perception-and-action physicality of human infants. Our simulation uses human data from videotaped infant/caregiver interactions and a rich 3D environment to model the development of gaze following. Initial tests suggest that infant gaze following can be learned in our simulation using parameters derived from behavioral data.

Keywords: embodiment; infancy; joint attention; simulation; social learning.

Human communication is a dauntingly complex system to model. Consider a seemingly simple system like an infant and caregiver playing together: even with language pared away, infant/caregiver social interactions feature a wide range of behaviors. These take place across many time scales in a complex environment. Moreover, the infant is a moving target; its brain and behavior change rapidly, and this requires caregivers to adapt to the infant's changing skills. Thus it is difficult to generate a powerful model of infant social behavior and learning.

Developing such a model is important because there is ample evidence that early social development has long term effects on (and likely serves as a foundation for) later social cognition, language, and even cognitive style and exploratory behavior [1]. In this paper we describe a modeling approach that is unique in two key areas, extending the approach introduced in [2]. First, we model both the learning agent (in this case the infant) and the agent's environment. Many models of infant learning use an abstract symbolic environment with little relation to the dynamic world infants experience. Ideally, simulations are comprised of both a biologically plausible learning model, and a physically and socially realistic environment [3]. The latter requirement is problematic because detailed data on the structure of infants' learning environment only exist in bits and pieces. Our second innovation is to directly tie behavioral data collected by our lab into our simulation environment, creating rich and realistic stimuli for our learning agent.

In the following subsections we will review the theoretical issues relevant to this work.

Embodied Modeling The goal of developmental modeling is to test theories of learning processes as they take place within organisms undergoing gross changes. Valid tests of these theories require additional theories as to the information patterns found in realistically structured environments [3]. Currently, however, we do not possess the computational resources to model human perceptual and neural systems, and our technological ability to simulate real, multi-modal environments is still primitive. The key, then, is to gradually converge on a set of biological traits that capture key properties of learning, as well as some key ecological patterns that can be simulated at a level of detail that is appropriate for the theoretical question at hand. This typically requires consideration of the physicality of the organism and the environment. That is, to test our theories with greater validity we must incorporate the embodiment of our models [4]. To the degree that we can embody simulations, we improve our tests of the motivating theory of development and learning.

Robotic studies are one way to achieve embodied simulations, and there are a growing number of good examples [5, 6, 7]. Robots can be placed in the same environments as infants and presented with identical stimuli. Unfortunately robotic studies are expensive, and they introduce tangential methodological issues—they require solving mechanical and computational problems simply to begin testing learning theories. Solving these problems is certainly important for some theoretical questions, but it is not currently necessary to address basic questions about infant social development. Additionally, robotic models cannot be run faster than real time, and they require active supervision. In many cases, current theories can realize faster progress by using simulations that retain elements of embodiment while greatly simplifying implementation and reducing cost.

Gaze Following We have been investigating the development of attention sharing behaviors in human infants. Attention sharing is a behavioral cornerstone of all social learning. In general it means one or more agents changing their focus of attention because they have observed another individual attending to some stimulus or area. A common example is following the line-of-gaze of another person. There is an extensive literature on the development of infants' attention sharing skills. This literature has focused on the development of gaze following, which is defined as reorienting one's direction of gaze to intersect with that of another person, based on encoding the other's head pose and/or eye direction.

Infants begin following other people's gaze between 6 and 12 months of age, and their ability to follow more and more subtle cues, to a wider range of their environment, increases significantly between 9 and 18 months of age [8, 9]. It is unknown by what mechanism infants develop more powerful gaze-following skills.

We have hypothesized [10] that infants' gaze following skills might emerge as the byproduct of a "basic set" of perceptual, learning, and affective traits that are in place within the first 2 to 3 months of age, well before fully developed gaze following can be observed. The basic set theory states that the following elements are sufficient (though not necessary) for joint attention:

- A set of motivational biases, in particular a preference for social stimuli such as human faces.
- Habituation as a basic reward attenuation mechanism.
- A learning mechanism such as temporal difference learning [11], to learn the temporal structure of predictable, contingent interactions between infant and caregiver.
- Early emerging perceptual traits such as attention shifting, face processing and sensitivity to motion, contrast, and color.
- A structured environment providing strong correlation between where caregivers look and where interesting things are.

This basic set of infant traits might be sufficient to generate new attention sharing skills. However, this requires that the infant learn on a regular regimen of well structured social input, as provided by an organized caregiver [10]. Our modeling efforts are meant to prove the plausibility of this theory. If they are unsuccessful, then perhaps additional mechanisms, such as special-purpose modules, are necessary for an agent to learn gaze following skills during the first 6-9 months of human social experience. The question, then, is how to generate valid simulations of this social learning process. We must imbue the simulated infant with biologically plausible perceptual, learning, and motivational traits, and we must imbue its environment with a reasonable facsimile of a natural social environment.

Naturalistic Social Coding The fine-grained structure of infant social environments is difficult to quantify. Although it is possible to derive gross patterns from previous observational and ethnographic behavioral studies, these tend to be sparse in details, and coded at such a low sampling rate that there is no information about caregivers' meaningful moment-by-moment action patterns. In most experimental studies of infant social responses, the social input from the adult is controlled and extremely artificial (e.g. [9]). Although these experimental studies are critical for establishing developmental "benchmarks" that a simulated infant should replicate, they do not provide information about real infant learning environments, which can be abstracted for simulation.

Our approach to solving this problems starts by generating a dense, rich video dataset of minimally directed interactions between infants and caregivers. Figure 1 shows one frame of these interactions from two separate viewpoints. By coding these interactions at 30fps in the manner described below, we generate a temporally detailed dataset that opens a new window into infant/caregiver interaction in a natural setting.

In the following sections we will explain our methodological workflow, describe the machine learning and computer vision techniques driving our simulated infant, present results from the simulation environment, and finally discuss the impact this work has on the modeling of infant social interaction.

Workflow

Our lab takes an end-to-end approach to infant social modeling (see Figure 2)—we start in the lab and in the homes of our subjects by collecting hours of audiovisual data from infant/caregiver interactions. These data consist of both seminaturalistic free play sessions and scripted lab sessions. In the free play sessions caregivers are instructed to play with their infants using a supplied set of toys while the infant is seated in a tray chair. In lab sessions an experimenter performs a series of gaze and point maneuvers to salient objects in the room while holding the infant's attention. In both cases the interactions are recorded with audio from multiple camera angles. The lab has amassed many terabytes of this audiovisual data, which is passed off to a team of undergraduate research assistants who perform a detailed frame by frame coding of relevant events (e.g. gaze shifts, manual actions, environmental and toy-generated noise). These codes are stored in a database in order to facilitate an automated analysis of infant/caregiver behavior using custom software written in C# and Python. The automated analysis derives information from the coding such as the probability of the infant or caregiver to transition from one state to another (e.g. from looking at a toy to looking at a social partner), the duration of their actions, and extended events where the infant and caregiver move through a specified series of states within a restricted time window [12].

Our simulation environment can operate in two modes. In the first, it simply replicates caregiver behavior from a particular experimental session using the codes in the database. If the real-life caregiver started off looking at the infant and then switched to looking at a toy after 2.3 seconds, the simulated caregiver will do the same. In the second mode, the care-

Figure 1: Still picture from naturalistic study, from which the simulated caregiver behavior is derived.

giver behaves probabilistically based on the transition probabilities and timings derived from the automated analysis. In this way, the caregiver behaves realistically without replicating the steps of any particular subject; the simulation can run indefinitely. For example, if our data indicate that caregivers transition from holding an object to holding and moving an object 20% of the time that they change what they are doing, then our simulation likewise will make that transition 20% of the time. In addition, this mode allows our caregiver to (in principle) respond contingently to previous actions of the infant. Our simulation environment is implemented in C++ and we use hardware-accelerated OpenGL for the 3D rendering. Unfortunately, to our knowledge there is no open software for human simulation, so we use Boston Dynamics' DI-Guy platform for rendering and animating our caregiver and props. Finally, at the end of the chain, our infant learning agent processes rendered frames of the simulation using the OpenCV computer vision library [13]. At each time step of the simulation the *only* information the infant agent receives about its environment are these rendered frames—it extracts a reward signal and high level information about the environment using the computer vision techniques described in the next section.

Methods

There are three primary components to our simulation, the caregiver and environment, the infant agent's visual processing system, and its learning system. In this section we will detail the three components, starting with the caregiver and environment.

Caregiver and Environment Our simulation environment is set in the interior of a room containing a table and a chair. The caregiver is seated at the chair and interacts with toys placed on the table (see Figure 3, top). The caregiver is capable of interacting with more than one toy, but for our initial simulations we used just one toy, a red bus, for simplicity. The simulated caregiver occupies several different attention and action states. It can be: waving or not waving its arm, looking at the infant or the toy, and holding the toy or not. These states correspond to codes for caregiver motion, caregiver gaze target, and caregiver held object status in our empirical data. Because our caregiver is simulated as an actual body, these discrete behavior states manifest to the infant as a wide range of visual stimuli. For example while waving an object the caregiver's arm can be in many positions. Similarly, when looking to an object the caregiver's head pose varies over time as the motion is undertaken and the final head pose is based on the actual position of the object in the room.

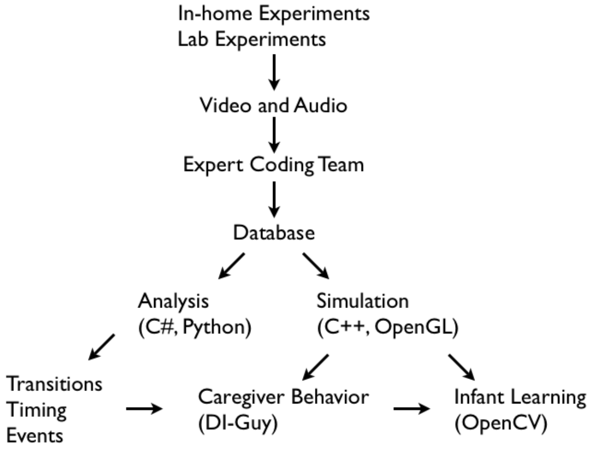

Figure 2: A flow chart depiction of the data collection, analysis and modeling work in our lab, annotated with relevant technologies.

From these data we also estimate the probability of transitioning between any of the states, and the simulated caregiver chooses its actions probabilistically based on these estimates (the caregiver is operating in the second mode described above, not off a script). The caregiver uses two transition matrices: the first governs behavior with respect to the toy (holding and waving) and the second governs looking target. The interval between state transitions is based on the observed interval between separate caregiver behaviors (every 2.18 seconds) plus some uniform noise (+/- 1 second).

The infant also has a body in the environment (unseen from its perspective), with its head at about high-chair height. Changes in infant gaze target are accomplished by tying the position and orientation of a camera to the position and orientation of this body's head.

The objects in the environment are part of the DI-Guy package, which has a nice variety of (mostly military themed) props. A text configuration file specifies the props to load at the start of the simulation as well as their location, orientation and scale. Similarly, the text file specifies the initial location, orientation and appearance of human agents. In this way we can quickly modify the appearance of the simulation, add agents, and rearrange props.

Visual Processing In order for the infant agent to learn from its raw visual input , it needs to extract high level information about its environmental state as well as determine the reward value of the state that it is in. Since we are interested in gaze following, we extract the caregiver head position from the raw image, estimate head pose and use the discretized pose state as the infant agent's environmental state. To do this we first localize the caregiver's head by calculating the probability that each pixel in the raw image came from the known distribution of pixel properties in caregiver head pixels, running a Gaussian blur over that probability map, and then centering a head position rectangle over the maximum probability point on the blurred map. Technically, this is an application of `cvCalcBackProject` (to calculate the back projection of our face color histogram), `cvSmooth` (the Gaussian blur) and `cvMinMaxLoc` (to find the location of maximum probability in the image) from the OpenCV library. Pragmatically, we're only assuming the infant knows broadly what its caregiver's face looks like.

To calculate the head pose, we break the detected head region up into a left and a right segment then perform a color histogram comparison between the observed segments and model segments of left and right facing heads (using `cvCompareHist`). From the histogram distances we can calculate the probability that the caregiver is looking left or right by seeing how close the observed segments are to the models. If the segments are distant from both models we can infer that the caregiver's head pose is center. Again, the only assumption is that the infant knows generally what left and right facing heads look like. Finally we discretize the head pose probability into three states: left, center, and right. A visualization of this head position and pose detection can be seen in Figure 3, middle. The box represents the head position and the handles represent the pose probability.

To compute the reward for the current frame of input, we first calculate a salience map over the entire frame. The salience map has three components: motion, contrast, and saturation, and it is similar to salience-based visual processing approaches such as the one in [6]. The components are summed to represent overall saliency. Motion is calculated by comparison with the previous input frame (`cvAbsDiff`), contrast is derived from an edge detection routine (`cvSobel`), and saturation is extracted naturally

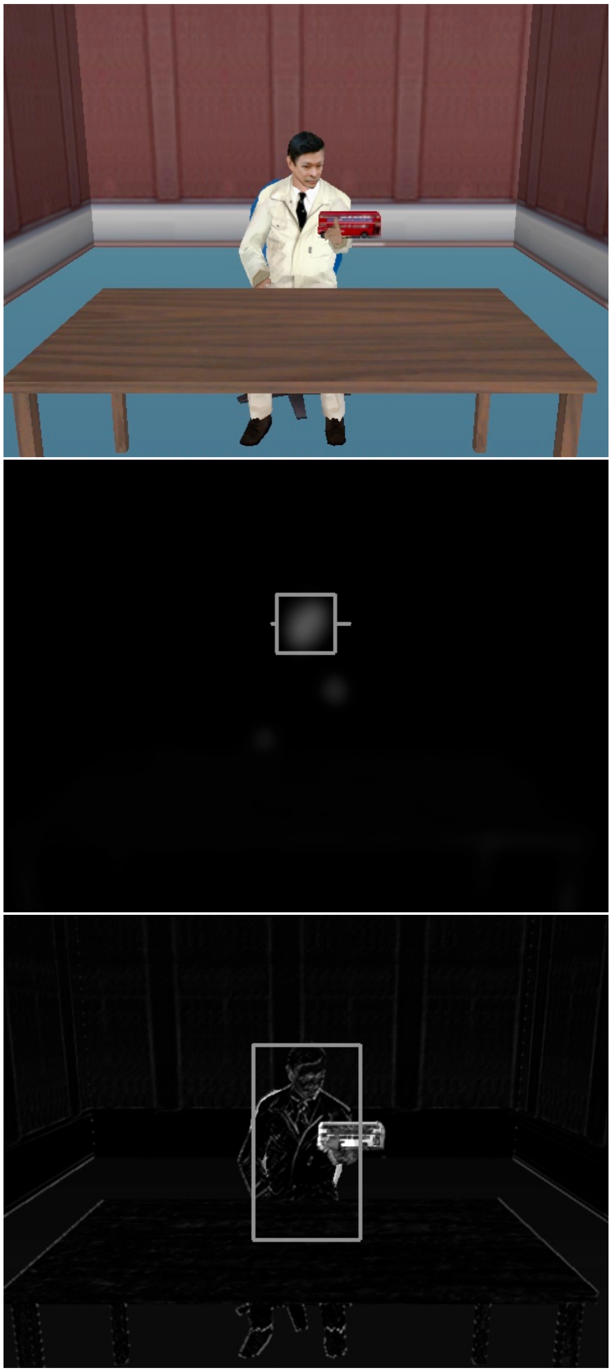

Figure 3: From top to bottom: the raw visual input to the infant agent, head detection and pose estimation output, salience and reward visualization.

from the color values of the pixels. Reward is then calculated by averaging the saliency values within the agent's center of vision (see Figure 3, bottom—the reward area is inside the rectangle). Since the agent only chooses looking direction on the horizontal axis, the center of vision is defined to be taller than it is wide.

Learning The agent uses a reinforcement learning [11] paradigm to choose its actions and learn from the consequences. Its state-action space is a cross of the discretized caregiver head poses and a set of five looking directions: left, near left, center, near right, and right. Every time the agent shifts gaze position, it updates its expected reward for the previous state-action pair using the following formula

$$er(i,j)_{new} = er(i,j)_{old} - \eta(er(i,j)_{old} - ar)$$

where $er(i,j)$ is the expected reward given caregiver head pose i and gaze action j, η is a learning rate parameter (set to 0.1 in our simulation) and ar is the average reward obtained since the last action j in state i. The agent changes gaze pose after a period of time derived from observed infant behavior (every 2.43 seconds) plus some uniform noise (+/- 1 second—a more complex but more realistic approach would be to draw fixation duration from an estimate of the fixation duration probability density function from actual infants).

It would be straightforward to increase the number of states and actions (e.g. by giving caregiver and infant looking direction a vertical degree of freedom) and add bells and whistles to the reward estimation process, but the purpose of this work is not to showcase machine learning techniques. Rather, we are investigating whether gaze following can be learned given a simple learning mechanism, data-driven caregiver behavior, and a complex simulated environment. The results of this endeavor are summarized and discussed in the next section.

Results

We ran our simulation for approximately 500 seconds (enough time for the infant to shift gaze about 200 times) with the infant agent watching a simulated caregiver interact with a single toy. The agent's expected reward over its state-action space is detailed in the table below. Looking at a location in the room with background (i.e. smallest) saliency results in a reward around 6.0, so that quantity is subtracted from the below numbers.

	Looking Direction				
CG Pose	left	near left	center	near right	right
left	1.30	1.54	3.58	2.62	1.79
center	1.09	2.62	8.50	3.20	1.97
right	1.56	1.72	1.71	1.43	0.76

Table 1: The final state/action reward space of the infant learning agent.

The course of learning over time is shown in Figure 4. The agent quickly learns that congruent gaze shifts result in higher reward and the advantage in expected reward generally increases over time.

Discussion

After a fairly short period of training, the agent expects more reward when its looking direction is congruent with the caregiver's head pose than when its looking direction is incongruent. For example, if the caregiver is looking to its left, then if the infant looks to the right it expects more reward (the infant and caregiver are facing each other and thus their looking directions to the same location are opposite). Both the near and far looking directions show this effect. Looking right in general is privileged because the caregiver is left handed (it only picks up objects with its left hand), and thus during time periods where the caregiver is holding the toy it is more likely to be near or far right than near or far left (from the infant's perspective).

Looking center is always very rewarding since the caregiver is at center. When the caregiver holds an object it will often be at center, and when it moves the object it generally is at center or near right. Motion is highly rewarding, and the caregiver is normally looking at center during motion, so the center/center expected reward is quite high. The caregiver also has a naturally higher contrast than other parts of the environment.

This general pattern of results fits other recent findings. It seems that infants in everyday setting are highly attentive to caregivers' manual actions [12], and this might bootstrap infants' learning of caregivers' head pose (because adults often look at what they are manipulating). It is also known that infants are attracted to faces, and the simulation results are consistent with that. Since the head pose and position estimation are not used in calculating reward, the infant agent learns that looking center (where the caregiver's head is) is valuable independent of the general knowledge about head appearance that it has.

These first results show that with a limited set of assumptions, a simple learning model, and a complex data-driven environment, gaze following can be learned. More importantly, this work sets the stage for even more detailed simulation of infant/caregiver interaction—such as interaction between more than two agents (a sibling agent, perhaps), reaching and grasping capability for the infants, and realistic audio cues. Further, since the infant agent no longer receives knowledge about its environmental state other than through visual processing, its input will degrade meaningfully and realistically. For example, if the infant picks up an object that occludes the caregiver, its head position and pose estimates will degrade realistically.

In the greater context of understanding infant social development, from modeling to robotics to experimental work, we see this as occupying a productive niche between disembodied and discretized 2D models and robotic agents. We open computer simulations up to state and action space complexities that mirror those in the real world, but our learning

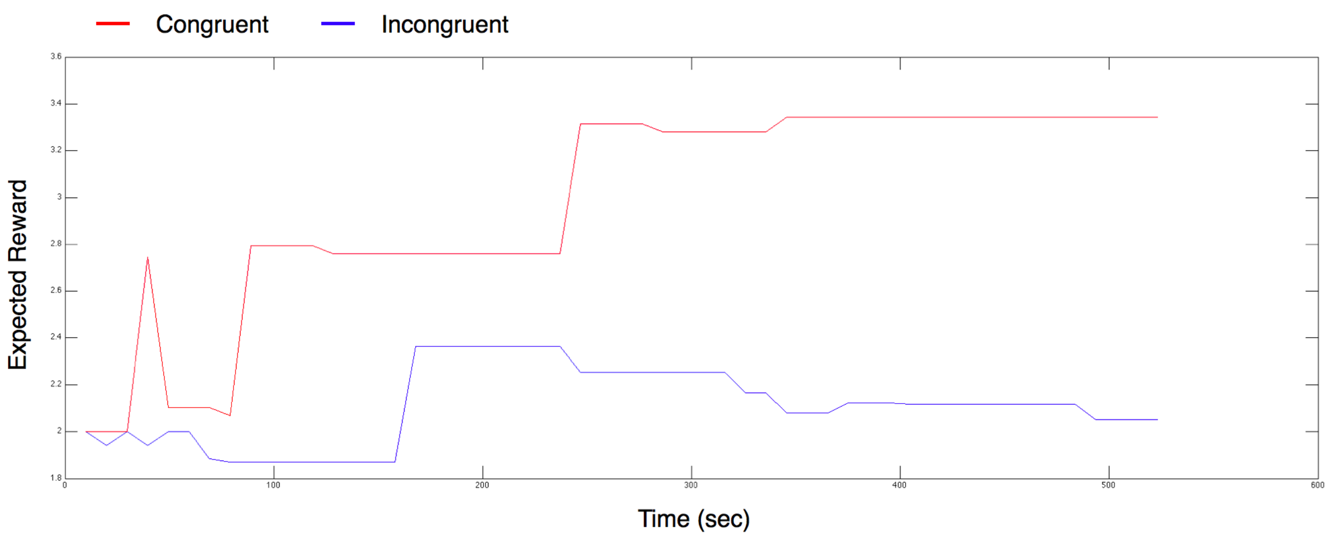

Figure 4: The sum reward expected from highly congruent gaze shifts (red, top right and bottom left of Table 1) and incongruent gaze shifts (blue, top left and bottom right) over the training period.

simulations are more convenient and we can have complete control over the agent and environment. Moreover, our simulations do not require the expensive and complicated hardware of robotic simulations; nor do they force us to address interesting but difficult and peripheral questions about motor control.

Acknowledgments

The authors would like to thank Ricky Ng for his valuable assistance on this project. This work was supported by research grants from the M.I.N.D. Institute and the National Alliance for Autism Research to G. Deák and J. Triesch, a National Science Foundation (SES-0527756) to G. Deák, and NSF IGERT Grant #DGE-0333451 to G. Cottrell and V de Sa.

References

[1] L. A. Sroufe, B. Egeland, E. Carlson, and W.A. Collins. *The Development of the Person: The Minnesota Study of Risk and Adaptation from Birth to Adulthood.* Guilford, New York, 2005.

[2] H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In *Biomimetic Neural Learning for Intelligent Robots*, volume 3575/2005, pages 211–224. Springer Berlin / Heidelberg, 2005.

[3] G. O. Deák, M.S. Bartlett, and T. Jebara. How social agents develop: New trends in integrative theory-building. *Neurocomputing*, 70:2139–2147, 2007.

[4] M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9:625636, 2002.

[5] G. Metta, G. Sandini, G. S., and L. Natale. Sensorimotor interaction in a developing robot. In *First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 18–19. Lund University Press, 2001.

[6] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 20:211–229, 2003.

[7] N. Butko, I. Fasel, and J. R. Movellan. Learning about humans during the first 6 minutes of life. *Proceedings of the International Conference on Development and Learning*, 2006.

[8] G. Butterworth and N. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:5572, 1991.

[9] G. O. Deák, R. A. Flom, and A. D. Pick. Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, 36:157–192, 2000.

[10] J. Triesch, C. Teuscher, G. O. Deák, and E. Carlson. Gaze-following: why (not) learn it? *Developmental Science*, 9:125–147, 2006.

[11] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[12] A. Krasno, G. Deák, J. Triesch, and H. Jasso. Watch the hands: Do infants learn gaze-following from parents' object manipulation? In *Biennial Meeting of the Society for Research in Child Development*, 2007.

[13] OpenCV Wiki. `http://opencv.willowgarage.com/wiki/`.