Imitation and Mechanisms of Shared Attention: A Developmental Structure for Building Social Skills

Brian Scassellati*
MIT Artificial Intelligence Lab
545 Technology Square
Cambridge, MA, 02139, USA
scaz@ai.mit.edu

Abstract

This paper explores the use of imitation in an on-going research program aimed at enabling a humanoid robot to communicate naturally with humans. Our group has constructed an upper-torso humanoid robot, called Cog, in part to investigate how to build intelligent robotic systems by following a developmental progression of skills similar to that observed in human development. Just as a child learns social skills and conventions through interactions with its parents, our robot will learn to interact with people using natural social communication. Our models of social interaction are drawn from developmental models of normal children, developmental models of autism, and on models of the evolutionary development of social skills.

In this paper, we consider the role that imitation plays in the development of a critical pre-cursor of normal human social development, mechanisms of shared attention. Mechanisms of shared attention serve to direct two individuals to attend to the same object in the environment, through eye direction, pointing gestures, and other means. Imitation serves a critical role in bootstrapping a system from simple eye behaviors to more complex social skills. We will present data from a face and eye finding system that serves as the basis of this developmental chain, and a short example of how this system can imitate the head movements of an individual.

1 Motivation

While the past few decades have seen increasingly complex machine learning systems, the systems we have constructed have failed to approach the flexibility, robustness, and versatility that humans display. There have been successful systems for extracting environmental invariants and exploring static environments, but there have been few attempts at building systems that learn by interacting with people using natural, social cues. With the advances of embodied systems research, we can now build systems that are robust enough, safe enough, and stable enough to allow machines to interact with humans in a learning environment.

One of the critical precursors to social learning in human development is the ability to selectively attend to an object of mutual interest. Humans have a large repertoire of social cues, such as gaze direction, pointing gestures, and postural cues, that all indicate to an observer which object is currently under consideration. These abilities, collectively named mechanisms of shared (or joint) attention, are vital to the normal development of social skills in children (Scaife & Bruner 1975). The primary focus of the research reported here is to investigate how individuals develop the skills to recognize and produce these social cues by implementing models of this developmental progression on a humanoid robot (see Figure 1). A more detailed account of this project can be found in (Scassellati 1996).

We are interested in shared attention as a precursor to social communication for two reasons. First, we believe that by using a developmental program to build social capabilities we will be able to achieve a wide range of natural interactions with untrained observers (Brooks, Ferrell, Irie, Kemp, Marjanovic, Scassellati & Williamson 1998). Constructing a machine that can recognize the social cues from a human observer allows for more natural human-machine interaction and creates possibilities for machines to learn by directly observing untrained human instructors. Second, by building models from developmental psychology and from studies of autism, we further these models by providing a test bed for manipulating the behavioral progression. With an implemented developmental model, we can test alternative learning conditions, environmental conditions, and evaluate alternative intervention and teaching techniques. This investigation of shared attention asks questions about the development and origins of the complex non-verbal communication skills that humans so easily master: What is the progression of skills that humans must acquire to engage in shared attention? When something goes wrong in this development, as it seems to do in autism, what problems can occur, and what hope do we have for correcting these problems? What parts of this complex interplay can be seen in other primates, and what can we learn about the basis of communication from these comparisons?

*The author receives support from a National Defense Science and Engineering Graduate Fellowship. Support for this project is provided in part by an ONR/ARPA Vision MURI Grant (No. N00014-95-1-0600).

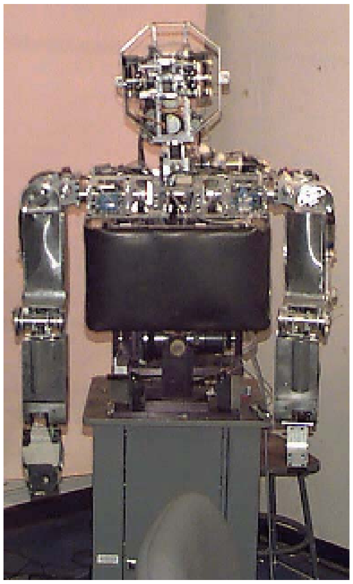

Figure 1: Cog, an upper-torso humanoid robot. Cog has twenty-one degrees of freedom to approximate human movement, and a variety of sensory systems that approximate human senses, including visual, vestibular, auditory, and tactile senses.

2 A Developmental Model of Shared Attention

Rather than appealing to an introspective description of social skills, and rather than implementing an ad-hoc hierarchy of skills, we instead rely upon studies from developmental psychology, abnormal psychology, and evolutionary psychology to provide insight into how nature has solved the problems of social communication. By studying the way that nature has decomposed this task, we hope not only to find ways of breaking our computational problem into manageable pieces, but also to explore some of the theories of human development.

The most relevant studies to our purposes have occured as developmental and evolutionary investigations of "theory of mind" (see Whiten (1991) for a collection of these studies). The most important finding, repeated in many different forms, is that the mechanisms of shared attention are not a single monolithic system. Evidence from childhood development shows that not all mechanisms for shared attention are present from birth, and there is a stereotypic progression of skills that occurs in all infants at roughly the same rate (Hobson 1993). There are also developmental disorders, such as autism, that limit and fracture the components of this system (Frith 1990). Additionally, this same ontological progression can be seen as an evolutionary progression in which the increasingly complex set of skills can be mapped to animals that are increasingly closer to humans on a phylogenetic scale (Povinelli & Preuss 1995).

As the basis for our implementation of shared attention, we turn to a developmental model from Baron-Cohen (1995). Baron-Cohen's model gives a coherent account of the observed developmental stages of shared attention behaviors in both normal and blind children, the observed deficiencies in shared attention of autistic children,[1] and a partial explanation of the observed abilities of primates on shared attention tasks. Baron-Cohen describes four Fodorian modules: the eye-direction detector (EDD), the intentionality detector (ID), the shared attention module (SAM), and the theory-of-mind module (TOMM). In brief, the eye-direction detector locates eye-like shapes and extrapolates the object that they are focused upon while the intentionality detector attributes desires and goals to objects that appear to move under their own volition. The outputs of these two modules (EDD and ID) are used by the shared attention module to generate representations and behaviors that link attentional states in the observer to attentional states in the observed. Finally, the theory-of-mind module acts on the output of SAM to predict the thoughts and actions of the observed individual.

Baron-Cohen's model gives a theoretical framework that accounts for both normal and abnormal development. What the model does not provide is a task-level decomposition of necessary skills and the developmental mechanisms that provide for transition between his stages. Our current work is on identifying and implementing a developmental account of one possible skill decomposition, an account which relies heavily upon imitation. The skill decomposition that we are pursuing can be broken down into four stages: maintaining eye contact, deictic gaze following, imperative pointing, and declarative pointing. In terms of Baron-Cohen's model, we are implementing a vertical slice of behaviors from parts of EDD, ID, and SAM that additionally matches the observed phylogeny of these skills.

The first step in producing mechanisms of shared attention is the recognition and maintenance of eye contact. Many animals have been shown to be extremely sensitive to eyes that are directed at them, including reptiles like the hognosed snake (Burghardt & Greene 1990), avians like the chicken (Scaife 1976) and the plover (Ristau 1991), and all primates (Cheney & Seyfarth 1990). Identifying whether or not something is looking at you provides an obvious evolutionary advantage in escaping predators, but in many mammals, especially primates, the recognition that another is looking at you carries social significance. In monkeys, eye contact is significant for maintaining a social dominance hierarchy (Cheney & Seyfarth 1990). In humans, the reliance on eye contact as a social cue is even more striking. Infants have an innate preference for looking at human faces and eyes, and maintain (and thus recognize) eye contact within the first three months. Maintenance of eye contact will be the behavioral goal for a system in this stage.

The second step is to engage in shared attention through deictic gaze following. Deictic gaze recognition is the capability to look not at an individual but at what that individual is attending to. Gaze following is the rapid alternation between looking at the eyes of the individual and looking at the distal object. While many animals are sensitive to eyes that are gazing directly at them, only great apes[2] show the capability to extrapolate from the direction of gaze to a distal object (Povinelli & Preuss 1995). This evolutionary progression is also mirrored in the ontogeny of social skills. At least by the age of three months, human infants display maintenance (and thus recognition) of eye contact. However, it is not until nine months that children gener-

[1] While the deficits of autism certainly cover many cognitive abilities, some researchers believe that the missing mechanisms of shared attention may be critical to the other deficiencies (Baron-Cohen 1995). In comparison to other mental retardation and developmental disorders (like Williams and Downs Syndromes), the deficiencies of autism in this area are quite specific (Karmiloff-Smith, Klima, Bellugi, Grant & Baron-Cohen 1995).

[2] The terms "monkey" and "ape" are not to be used interchangeably. Apes include orangutans, gorillas, bonobos, chimpanzees, and humans. All apes are monkeys, but not all monkeys are apes.

ally exhibit deictic gaze following, and not until eighteen months that children will follow gaze outside their field of view (Baron-Cohen 1995). Deictic gaze following is an extremely useful imitative gesture which serves to focus the child's attention on the same object that the caregiver is attending to. This functional imitation appears simple, but a complete implementation of deictic gaze following involves many separate proficiencies, as we will discuss in the following section.

The third step in our account is imperative pointing. Imperative pointing is a gesture used to request an object that is out of reach by pointing at that object. This behavior is first seen in human children at about nine months of age (Baron-Cohen 1995), and occurs in many monkeys (Cheney & Seyfarth 1990). From the child's perspective, imperative pointing is a relatively simple extension of normal reaching behavior. One can imagine the child learning this behavior through simple reinforcement; the reaching motion of the infant is interpreted as a request by the adult for a specific object, which the adult then acquires and provides to the child. There is nothing particular to the infant's behavior that is different from a simple reach – it is the interpretation of the caregiver that provides meaning. Generation of this behavior is then a simple extension of primitive reaching behavior.

The fourth step is the advent of declarative pointing. Declarative pointing is characterized by an extended arm and index finger designed to draw attention to a distal object. Unlike imperative pointing, it is not necessarily a request for an object, and thus requires more complex computational mechanisms. We propose that imitation is a critical factor in the ontogeny of declarative pointing. This is an appealing speculation from both an ontological and a phylogenetic standpoint. From an ontological perspective, declarative pointing begins to emerge at approximately 12 months in human infants, which is also the same time that other complex imitative behaviors such as pretend play begin to emerge. From the phylogenetic perspective, declarative pointing has not been identified in any non-human primate (Premack 1988). This also corresponds to the phylogeny of imitation; no non-human primate has ever been documented to display true imitative behavior (Hauser 1996). We propose that the child first learns to recognize the declarative pointing gestures of the adult and then imitates those gestures in order to produce declarative pointing. The recognition of pointing gestures builds upon the competencies of gaze following; the infrastructure for extrapolation from a body cue is already present from the first two stages, it need only be applied to a new domain. The generation of declarative pointing gestures requires the same motor capabilities as imperative pointing, but it must be utilized in specific social circumstances. By imitating the successful pointing gestures of other individuals, the robot can learn to make use of similar gestures.

3 Current Results

In the past two years, we have focused on developing the sensori-motor coordination and basic perceptual capabilities for our humanoid robot. With a basic repertoire of sensorimotor and perceptual skills, we can begin to construct the developmental program outlined above. The hardware platform that we use for vision is a binocular, foveated, active vision system designed to mimic some of the capabilities of the human visual system (Scassellati 1998*a*).[3] To allow for both a wide field of view and high resolution vision, there are two cameras per eye, one which captures a wide-angle view of the periphery (approximately 110° field of view) and one which captures a narrow-angle view of the central (foveal) area (approximately 20° field of view with the same resolution). The robot also has a three degree of freedom neck and a pair of human-like arms. Each arm has six compliant degrees of freedom, each of which is powered by a series elastic actuator (Pratt & Williamson 1995) which provides a sensible "natural" behavior: if it is disturbed, or hits an obstacle, the arm simply deflects out of the way. Additional details of the hardware implementation can be found in Brooks & Stein (1994).

3.1 Implementing Maintenance of Eye Contact

Implementing the first stage in our developmental framework, recognizing and responding to eye contact, requires mostly perceptual abilities. We require that the robot be capable of finding faces, of determining the location of the eye within the face, and of determining if the eye is looking at the robot. Many computational methods of face detection on static images have been investigated by the machine vision community, for example (Sung & Poggio 1994, Rowley, Baluja & Kanade 1995). However, these methods are computationally intensive, and current implementations do not operate in real time. However, a simpler strategy for finding faces can operate in real time and produce reasonably good results.

The strategy that we use is based on the ratio-template method of object detection reported by Sinha (1994). In summary, finding a face is accomplished with the following five steps:

1. Use a motion-based pre-filter to identify potential face locations in the peripheral image
2. Use a ratio-template based face detector to identify target faces
3. Saccade to the target using a learned sensori-motor mapping
4. Convert the location in the peripheral image to a foveal location using a learned mapping
5. Extract the image of the eye from the foveal image

A short summary of these steps appears below, and additional details can be found in Scassellati (1998*b*).

To identify face locations, the peripheral image is converted to grayscale and passed through a pre-filter stage. The pre-filter allows us to search only locations that are likely to contain a face, greatly improving the speed of the detection step. The pre-filter selects a location as a potential target if it has had motion in the last 4 frames, was a detected face in the last 5 frames, or has not been evaluated in 3 seconds. A combination of the pre-filter and some early-rejection optimizations allows us to detect faces at 20 Hz with little accuracy loss.

Face detection is done with a method called "ratio templates" designed to recognize frontal views of faces under

[3]Two additional copies of this platform exist as desktop development platforms. While there are minor differences between the platforms, these differences are not important to the work reported here. Some of the results in this section were obtained from those platforms.

varying lighting conditions (Sinha 1996). A ratio template is composed of a number of regions and a number of relations, as shown in Figure 2. Overlaying the template with a grayscale image location, each region is convolved with the grayscale image to give the average grayscale value for that region. Relations are comparisons between region values, such as "the left forehead is brighter than the left temple." In Figure 2, each arrow indicates a relation, with the head of the arrow denoting the lesser value. The match metric is the number of satisfied relations; the more matches, the higher the probability of a face.

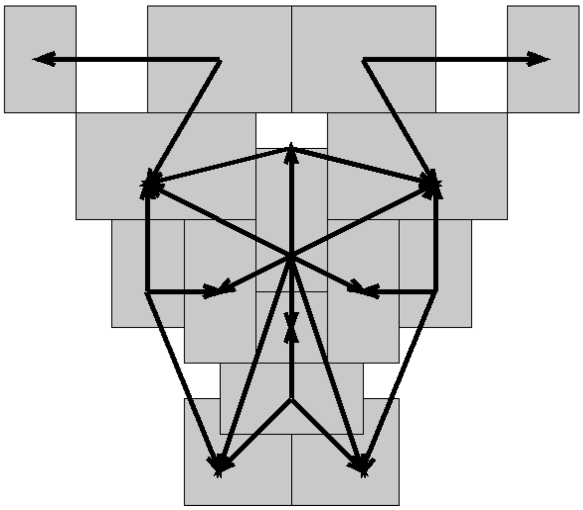

Figure 2: A ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows).

Once a face has been detected, the face location is converted into a motor command to center the face in the peripheral image using a learned saccade map. This map is implemented as a 17×17 interpolated lookup table, which is trained by the following algorithm:

1. Initialize with a linear map obtained from self-calibration
2. Randomly select a visual target
3. Saccade using the current map
4. Find the target in the post-saccade image using correlation
5. Update the saccade map based on L_2 error
6. Go to step 2

More information on this technique can be found in Marjanović, Scassellati & Williamson (1996). The system converges to an average of less than one pixel of error per saccade after 2000 trials (1.5 hours).

After the active vision system has saccaded to the face, the face and eye locations from the template in the peripheral camera are mapped into the foveal camera using a second learned mapping. The mapping from foveal to peripheral pixel locations can be seen as an attempt to find both the difference in scales between the images and the difference in pixel offset. In other words, we need to estimate four parameters: the row and column scale factor that we must apply to the foveal image to match the scale of the peripheral image, and the row and column offset that must be applied to the foveal image within the peripheral image. This mapping can be learned in two steps. First, the scale factors are estimated using active vision techniques. While moving the motor at a constant speed, we measure the optic flow of both cameras. The ratio of the flow rates is the ratio of the image sizes. Second, we use correlation to find the offsets. The foveal image is scaled down by the discovered scale factors, and then correlated with the peripheral image to find the best match location.

Once this mapping has been learned, whenever a face is foveated we can extract the image of the eye from the foveal image. This extracted image is then ready for further processing. Figure 3 shows the result of the face detection routines on a typical grayscale image before the saccade. Figure 4 shows the extracted image of the eye that was obtained after saccading to the target face. Work on extracting the location of the pupil within the eye has begun, but is still in progress. In order to accurately recognize whether or not the caregiver is looking at the robot, we must take into account both the position of the eye within the head and the position of the head with respect to the body. Additional work on extracting the postural position of the head on the body has begun.

Figure 3: An example of the face detector. The 128x128 grayscale image was captured by the active vision system, and then processed by the pre-filtering and ratio template detection routines. One face was found within the image, and is shown outlined.

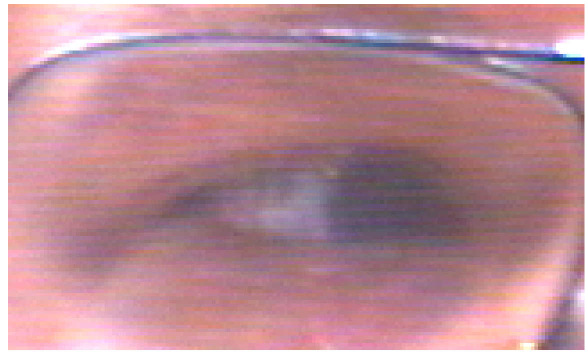

Figure 4: Extracted image of the eye from the foveal image.

3.2 Implementing Deictic Gaze Following

Once our system is capable of detecting eye contact, we require three additional subskills to achieve deictic gaze following: extracting the angle of gaze, extrapolating the angle of gaze to a distal object, and motor routines for alternating between the distal object and the caregiver. Extracting angle of gaze is a generalization of detecting someone gazing at you. Extrapolation of the angle of gaze can be more complex. By a geometric analysis of this task, we would need to determine not only the angle of gaze, but also the degree of vergence of the observer's eyes to find the distal object. However, the ontogeny of gaze following in human children points us to a somewhat simpler explanation. Butterworth (1991) has shown that at approximately 6 months, infants will begin to follow a caregiver's gaze to the correct side (left/right). Over the next three months, their accuracy increases so that they can determine the angle of gaze. At 9 months, the child will track from the caregiver's eyes along the angle of gaze until it encounters a salient object. Even if the actual object of attention is further along the angle of gaze, the child is somehow "stuck" on the first object encountered along that path. Butterworth labels this the "ecological" mechanism of joint visual attention, since it is the nature of the environment itself that completes the action. It is not until 12 months that the child will reliably attend to the distal object regardless of its order in the scan path. This "geometric" stage indicates that the infant successfully can determine not only the angle of gaze but also the vergence. However, at this stage, infants will only exhibit gaze following if there are objects within their field of view. They will not turn to look behind them, even if the angle of gaze from the caretaker would warrant. Around 18 months, the infant begins to enter a "representational" stage in which it will follow gaze angles outside its own field of view, that is, it somehow represents the angle of gaze and the presence of objects outside its own view.

Implementing this progression for a robotic system provides for a simple means of bootstrapping behaviors. The capabilities used in detecting and maintaining eye contact can be extended to provide a rough angle of gaze. By tracking along this angle of gaze, and watching for salient objects, we can match the ecological conditions. From an ecological mechanism, we can refine the algorithms for determining gaze, and add mechanisms for determining vergence. With feedback from the caregiver, this can be used to construct a geometric mechanism, which in turn can be generalized into a representational mechanism.

3.3 Implementing Imperative Pointing

Implementing imperative pointing is accomplished by implementing the more generic task of reaching to a visual target. Children pass through a developmental progression of reaching skills (Diamond 1990). The fist stage in this progression appears around the fifth month and is characterized by a very stereotyped reach which always initiates from a position close to the child's eyes and moves ballistically along an angle of gaze directly toward the target object. Should the infant miss with the first attempt, the arm is withdrawn to the starting position and the attempt is repeated.

To achieve this stage of reaching on our robotic system, we utilize the foveation behavior obtained from the first step in order to train the arm where to reach. To reach to a visual target, the robot must learn the mapping from camera image coordinates $\vec{x} = (x, y)$ to the head-centered coordinates of the eye motors $\vec{e} = (pan, tilt)$ and then to the coordinates of the arm motors $\vec{\alpha} = (\alpha_0 ... \alpha_5)$. The saccade map $\vec{S} : \vec{x} \rightarrow \vec{e}$ relates positions in the camera image with the motor commands necessary to foveate the eye at that location. Our task then becomes to learn the ballistic movement mapping from head-centered coordinates $\vec{e}$ to arm-centered coordinates $\vec{\alpha}$. To simplify the dimensionality problems involved in controlling a six degree-of-freedom arm, arm positions are specified as a linear combination of basis posture primitives. The ballistic mapping $\vec{B} : \vec{e} \rightarrow \vec{\alpha}$ is constructed by an on-line learning algorithm that compares motor command signals with visual motion feedback clues to localize the arm in visual space.

A single learning trial proceeds as follows:

1. Locate a visual target.
2. Saccade to that target using the learned saccade map.
3. Convert the eye position to a ballistic arm using the ballistic map.
4. Reach out the arm.
5. Use motion detection to locate the end of the arm.
6. Use the saccade map to convert the error signal from image coordinates into gaze positions, which can be used to train the ballistic map.
7. Withdraw the arm, and repeat.

This learning algorithm operates continually, in real time, and in an unstructured "real-world" environment without using explicit world coordinates or complex kinematics. This technique successfully trains a reaching behavior within approximately three hours of self-supervised training. Additional details on this method can be found in Marjanović et al. (1996).

3.4 Implementing Declarative Pointing

The task of recognizing a declarative pointing gesture can be seen as the application of the geometric and representational mechanisms for deictic gaze following to a new initial stimulus. Instead of extrapolating from the vector formed by the angle of gaze to achieve a distal object, we extrapolate the vector formed by the position of the arm with respect to the body. This requires a rudimentary gesture recognition system, but otherwise utilizes the same mechanisms.

We have proposed that producing declarative pointing gestures relies upon the imitation of declarative pointing in an appropriate social context. We have not yet begun to focus on the problems involved in recognizing these contexts, but we have begun to build systems capable of simple mimicry. By adding a tracking mechanism to the output of the face detector and then classifying these outputs, we have been able to have the system mimic yes/no head nods of the caregiver. As the caregiver shakes his head yes, the robot will also shake its head yes. While this is a very simple form of imitation, it is highly selective. Merely producing horizontal or vertical movement is not sufficient for the head to mimic the action – the movement must come from a face-like object. Video clips of this imitation are available from http://www.ai.mit.edu/projects/cog/Text/video-index.html.

References

Baron-Cohen, S. (1995), *Mindblindness*, MIT Press.

Brooks, R. & Stein, L. A. (1994), 'Building Brains for Bodies', *Autonomous Robots* **1:1**, 7–25.

Brooks, R. A., Ferrell, C., Irie, R., Kemp, C. C., Marjanovic, M., Scassellati, B. & Williamson, M. (1998), Alternative Essences of Intelligence, *in* 'Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)', AAAI Press.

Burghardt, G. M. & Greene, H. W. (1990), 'Predator Simulation and Duration of Death Feigning in Neonate Hognose Snakes', *Animal Behaviour* **36**(6), 1842–1843.

Butterworth, G. (1991), The Ontogeny and Phylogeny of Joint Visual Attention, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Cheney, D. L. & Seyfarth, R. M. (1990), *How Monkeys See the World*, University of Chicago Press.

Diamond, A. (1990), Developmental Time Course in Human Infants and Infant Monkeys, and the Neural Bases, of Inhibitory Control in Reaching, *in* 'Development and Neural Bases of Higher Cognitive Functions', Vol. 608, New York Academy of Sciences, pp. 637–676.

Frith, U. (1990), *Autism : Explaining the Enigma*, Basil Blackwell.

Hauser, M. D. (1996), *Evolution of Communication*, MIT Press.

Hobson, R. P. (1993), *Autism and the Development of Mind*, Erlbaum.

Karmiloff-Smith, A., Klima, E., Bellugi, U., Grant, J. & Baron-Cohen, S. (1995), 'Is there a social module? Language, face processing, and theory of mind in individuals with Williams Syndrome', *Journal of Cognitive Neuroscience* **7:2**, 196–208.

Marjanović, M., Scassellati, B. & Williamson, M. (1996), Self-Taught Visually-Guided Pointing for a Humanoid Robot, *in* 'From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB-96)', Bradford Books, pp. 35–44.

Povinelli, D. J. & Preuss, T. M. (1995), 'Theory of mind: evolutionary history of a cognitive specialization', *Trends in Neuroscience* **18**(9), 418–424.

Pratt, G. A. & Williamson, M. M. (1995), Series Elastic Actuators, *in* 'Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-95)', Vol. 1, Pittsburg, PA, pp. 399–406.

Premack, D. (1988), "Does the chimpanzee have a theory of mind?" revisited, *in* R. Byrne & A. Whiten, eds, 'Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.', Oxford University Press.

Ristau, C. A. (1991), Before Mindreading: Attention, Purposes and Deception in Birds?, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Rowley, H., Baluja, S. & Kanade, T. (1995), Human Face Detection in Visual Scenes, Technical Report CMU-CS-95-158, Carnegie Mellon University.

Scaife, M. (1976), 'The response to eye-like shapes by birds. II. The importance of staring, pairedness, and shape.', *Animal Behavior* **24**, 200–206.

Scaife, M. & Bruner, J. (1975), 'The capacity for joint visual attention in the infant.', *Nature* **253**, 265–266.

Scassellati, B. (1996), Mechanisms of Shared Attention for a Humanoid Robot, *in* 'Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium', AAAI Press.

Scassellati, B. (1998*a*), A Binocular, Foveated Active Vision System, Technical Report 1628, MIT Artificial Intelligence Lab Memo.

Scassellati, B. (1998*b*), Finding Eyes and Faces with a Foveated Vision System, *in* 'Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)', AAAI Press.

Sinha, P. (1994), 'Object Recognition via Image Invariants: A Case Study', *Investigative Ophthalmology and Visual Science* **35**, 1735–1740.

Sinha, P. (1996), Perceiving and recognizing three-dimensional forms, PhD thesis, Massachusetts Institute of Technology.

Sung, K.-K. & Poggio, T. (1994), Example-based Learning for View-based Human Face Detection, Technical Report 1521, MIT Artificial Intelligence Lab Memo.

Whiten, A., ed. (1991), *Natural Theories of Mind*, Blackwell.