Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot

Brian Scassellati

MIT Artificial Intelligence Lab
545 Technology Square
Cambridge MA 02139, USA
scaz@ai.mit.edu
http://www.ai.mit.edu/people/scaz/

Abstract. Adults are extremely adept at recognizing social cues, such as eye direction or pointing gestures, that establish the basis of joint attention. These skills serve as the developmental basis for more complex forms of metaphor and analogy by allowing an infant to ground shared experiences and by assisting in the development of more complex communication skills. In this chapter, we review some of the evidence for the developmental course of these joint attention skills from developmental psychology, from disorders of social development such as autism, and from the evolutionary development of these social skills. We also describe an on-going research program aimed at testing existing models of joint attention development by building a human-like robot which communicates naturally with humans using joint attention.
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. We further consider the critical role that imitation plays in bootstrapping a system from simple visual 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 an example of how this system can imitate the head movements of an individual.

1 Motivation

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 joint (or shared) attention, are vital to the normal development of social skills in children. Joint

C. Nehaniv (Ed.): Computation for Metaphors, Analogy, and Agents, LNCS 1562, pp. 176-195, 1999.

attention to objects and events in the world serves as the initial mechanism for infants to share experiences with others and to negotiate shared meanings. Joint attention is also a mechanism for allowing infants to leverage the skills and knowledge of an adult caretaker in order to learn about their environment, in part by allowing the infant to manipulate the behavior of the caretaker and in part by providing a basis for more complex forms of social communication such as language and gestures.

Joint attention has been investigated by researchers in a variety of fields. Experts in child development are interested in these skills as part of the normal developmental course that infants acquire extremely rapidly, and in a stereotyped sequence (Scaife & Bruner 1975, Moore & Dunham 1995). Additional work on the etiology and behavioral manifestations of developmental disorders such as autism and Asperger's syndrome have focused on disruptions to joint attention mechanisms and demonstrated how vital these skills are in our social world (Cohen & Volkmar 1997, Baron-Cohen 1995). Philosophers have been interested in joint attention both as an explanation for issues of contextual grounding and as a precursor to a theory of other minds (Whiten 1991, Dennett 1991). Evolutionary psychologists and primatologists have focused on the evolution of these simple social skills throughout the animal kingdom as a means of evaluating both the presence of theory of mind and as a measure of social functioning (Povinelli & Preuss 1995, Hauser 1996, Premack 1988).

We have approached joint attention from a slightly different perspective: the construction of human-like robots that exhibit these social skills (Scassellati 1996). This approach focuses first on the construction of useful real-world systems that can both recognize and produce normal human social cues, and second on the evaluation of the complex models of joint attention developed by other disciplines.

Building machines that can recognized human social cues will provide a flexibility and robustness that current systems lack. 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 advances in 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. 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. 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).

Robotics also offers a unique tool to developmental psychology and related disciplines in evaluating complex interaction models. By implementing these

models in a real-world system, we provide a test bed for manipulating the behavioral progression. With an implemented developmental model, we can test alternative learning and environmental conditions in order to evaluate alternative intervention and teaching techniques. This investigation of joint 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? With a robotic implementation of the theoretical models, we can further these investigations in previously unavailable directions.

However, building a robot with the complete social skills of a human is a Herculean task that still resides in the realm of science fiction and not artificial intelligence. In order to build a successful implementation, we must decompose the monolithic "social skills module" into manageable pieces. The remainder of this chapter will be devoted to building a rough consensus of evidence from work on autism and Asperger's syndrome, from developmental psychology, and from evolutionary studies on how this decomposition can best be accomplished. From this rough consensus, we will outline a program for building a robot that can recognize and generate simple joint attention behaviors. Finally, we will describe some of the preliminary steps we have taken with one humanoid robot to build this developmental program.

2 A Developmental Model of Joint Attention

To build complex social skills, we must have a decomposition of simpler behavioral skills that can be implemented and tested on our robotic system. This section will first describe why we believe that a decomposition is possible, based upon evidence from developmental psychology, abnormal psychology, and evolutionary psychology. 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. We then focus on one module-based decomposition of joint attention skills. With this as a theoretical basis, we then begin to develop a task-based decomposition which can be implemented and tested on a robotic system.

2.1 Evidence that Decomposition is Possible

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 joint attention are not a single monolithic system. Evidence from childhood development shows that not all mechanisms for

joint 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). For example, infants are always sensitive to eye direction before they can interpret and generate pointing gestures.

There are also developmental disorders, such as autism, that limit and fracture the components of this system (Frith 1990). Autism is a pervasive developmental disorder of unknown etiology that is diagnosed by a set of behavioral criteria centered around abnormal social and communicative skills (DSM 1994, ICD 1993). Individuals with autism tend to have normal sensory and motor skills, but have difficulty with certain socially relevant tasks. For example, autistic individuals fail to make appropriate eye contact, and while they can recognize where a person is looking, they often fail to grasp the implications of this information. While the deficits of autism certainly cover many other cognitive abilities, some researchers believe that the missing mechanisms of joint 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 social deficiencies of autism are quite specific (Karmiloff-Smith, Klima, Bellugi, Grant & Baron-Cohen 1995).

Evidence from research into the social skills of other animals has also indicated that joint attention can be decomposed into a set of subskills. The same ontological progression of joint attention skills that is evident in human infants can also 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). For example, skills that infants acquire early in life, such as sensitivity to eye direction, have been demonstrated in relatively simple vertebrates, such as snakes (Burghardt & Greene 1990), while skills that are acquired later tend to appear only in the primates (Whiten 1991).

2.2 A Module-Based Decomposition

As the basis for our implementation of joint attention, we begin with a developmental model from Baron-Cohen (1995). Baron-Cohen's model gives a coherent account of the observed developmental stages of joint attention behaviors in both normal and blind children, the observed deficiencies in joint attention of children with autism, and a partial explanation of the observed abilities of primates on joint 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.

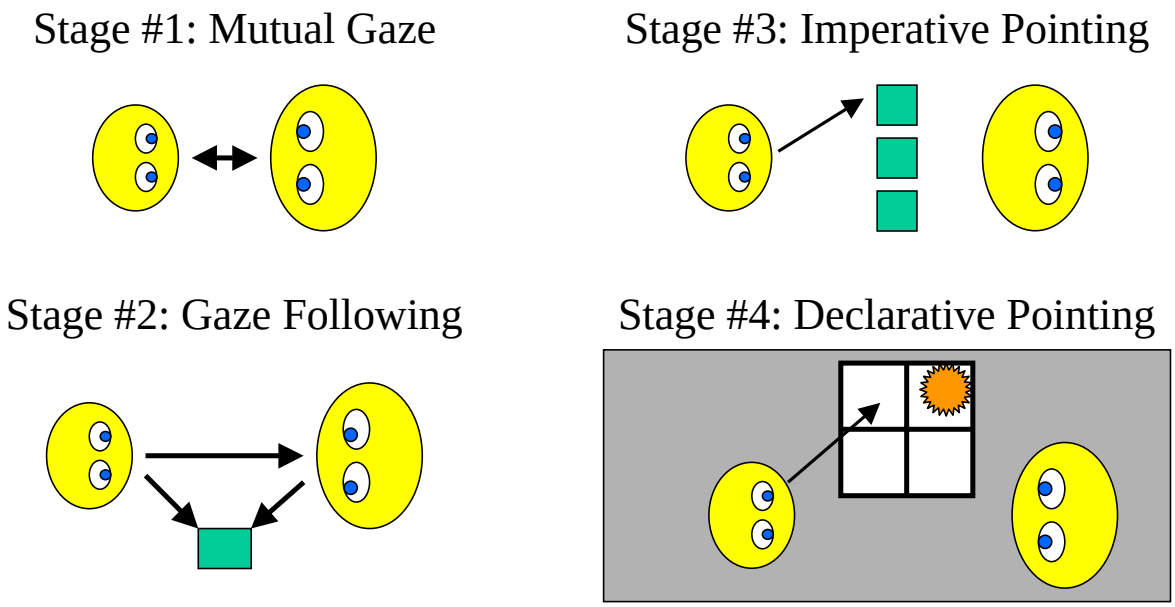

Fig. 1. A four-part task-based decomposition of joint attention skills. The capabilities for maintaining mutual gaze lead to the ability of gaze following. Imperative pointing skills, combined with gaze following, results in declarative pointing. For further information, see section 2.3.

This module-based description is a useful analysis tool, but does not provide sufficient detail for a robotic implementation. To build a portion of joint behavior skills, we require a set of observable behaviors that can be used to evaluate the performance of the system incrementally. We require a task-level decomposition of necessary skills and the developmental mechanisms that provide for transition between stages. Our current work is on identifying and implementing a developmental account of one possible skill decomposition, an account which relies heavily upon imitation.

2.3 A Task-Based Decomposition

The task-based skill decomposition that we are pursuing can be broken down into four stages: maintaining eye contact, gaze following, imperative pointing, and declarative pointing. Figure 1 shows simple cartoon illustrations of these four skills. The smaller figure on the left in each cartoon represents the novice and the larger figure on the right represents the caretaker. 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 joint 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 a strong 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 testable behavioral goal for a system in this stage.

The second step is to engage in joint attention through gaze following. Gaze following is the rapid alternation between looking at the eyes of the individual and looking at the distal object of their attention. While many animals are sensitive to eyes that are gazing directly at them, only primates show the capability to extrapolate from the direction of gaze to a distal object, and only the great apes will extrapolate to an object that is outside their immediate field of view (Povinelli & Preuss 1995).[1] 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 begin to exhibit gaze following, and not until eighteen months that children will follow gaze outside their field of view (Baron-Cohen 1995). 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 simplest form of joint attention is believed to be critical for social scaffolding(Thelen & Smith 1994), development of theory of mind(Baron-Cohen 1995), and providing shared meaning for learning language (Wood, Bruner & Ross 1976). This functional imitation appears simple, but a complete implementation of gaze following involves many separate proficiencies. Imitation is a developing research area in the computational sciences (for excellent examples, see (Dautenhahn 1994, Hayes & Demiris 1994, Dautenhahn 1997)).

The third step in our account is imperative pointing. Imperative pointing is a gesture used to obtain 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). However, there is nothing particular to the infant's behavior that is different from a simple reach – the infant is initially as likely to perform imperative pointing when the caretaker is attending to the infant as when the caretaker is looking in the other direction or when the caretaker is not present. The caregiver's interpretation of infant's gesture provides the shared meaning. Over time, the infant learns when the gesture is appropriate. One can imagine the child learning this behavior through simple reinforcement. The reaching motion of the infant is interpreted by the adult as a request for a specific object, which the adult then acquires

[1] 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.

and provides to the child. The acquisition of the desired object serves as positive reinforcement for the contextual setting that preceded the reward (the reaching action in the presence of the attentive caretaker). Generation of this behavior is then a simple extension of a 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; children often use declarative pointing to draw attention to objects that are clearly outside their reach, such as the sun or an airplane passing overhead. Declarative pointing also only occurs under specific social conditions; children do not point unless there is someone to observe their action. 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 imitative behavior under general conditions (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 and imperative pointing; the infrastructure for extrapolation from a body cue is already present from gaze following, 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 child can learn to make use of similar gestures.

3 Implementing Joint Attention

To build a system that can both recognize and produce the joint attention skills outlined above, we require a system with both human-like sensory systems and motor abilities. The Cog project at the MIT Artificial Intelligence Laboratory has been constructing 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 (Brooks & Stein 1994, Brooks et al. 1998). In the past two years, a basic repertoire of perceptual capabilities and sensory-motor skills have been implemented on the robot (see Brooks et al. (1998) for a review).

The 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. Cog's visual system is designed to mimic some of the capabilities of the human visual system,

Fig. 2. Images obtained from the peripheral (top) and foveal (bottom) cameras on Cog. The peripheral image is used for detecting salient objects worthy of visual attention, while the foveal image is used to obtain high resolution detail of those objects.

including binocularity and space-variant sensing (Scassellati 1998*a*). 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), as shown in Figure 2. Two additional copies of this active vision system are used as desktop development platforms, and were used to collect some of the data reported in the following sections. While there are minor differences between the platforms, these differences are not important to the work reported here. Cog 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.

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 at least that the robot be capable of (1) finding faces, (2) determining the location of the eye within the face, and (3) determining if the eye is looking at the robot. The only necessary motor abilities are to maintain a fixation point.

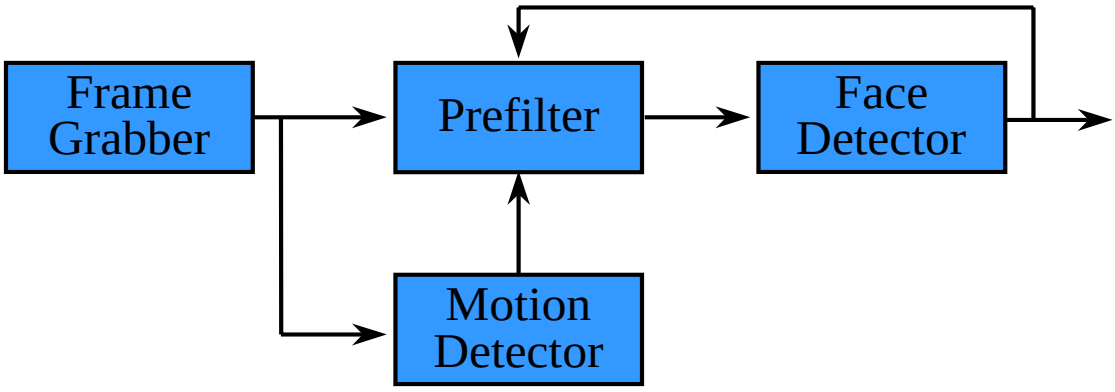

Fig. 3. Block diagram for the pre-filtering stage of face detection. The pre-filter selects target locations based upon motion information and past history. The pre-filter allows face detection to occur at 20 Hz with little accuracy loss.

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 good results under dynamic conditions (Scassellati 1998*b*). 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 sensory-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 (see Figure 3). 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 varying lighting conditions (Sinha 1996). A ratio template is composed of a number of regions and a number of relations,

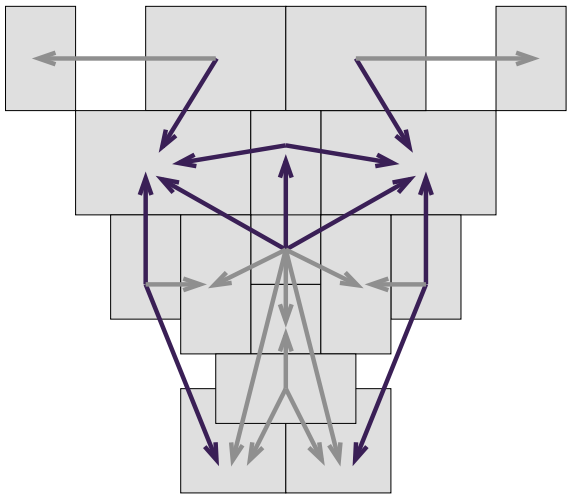

Fig. 4. A ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows).

as shown in Figure 4. 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 4, 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.

Once a face has been detected, the face location is converted into a motor command to center the face in the peripheral image. To maintain portability between the development platforms and to ensure accuracy in the sensory-motor behaviors, we require that all of our sensory-motor behaviors be learned by on-line adaptive algorithms (Brooks et al. 1998). The mapping between image locations and the motor commands necessary to foveate that target is called a 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.

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

Because humans are rarely motionless, after the active vision system has saccaded to the face, we first verify the location of the face in the peripheral image. The face and eye locations from the template in the peripheral camera

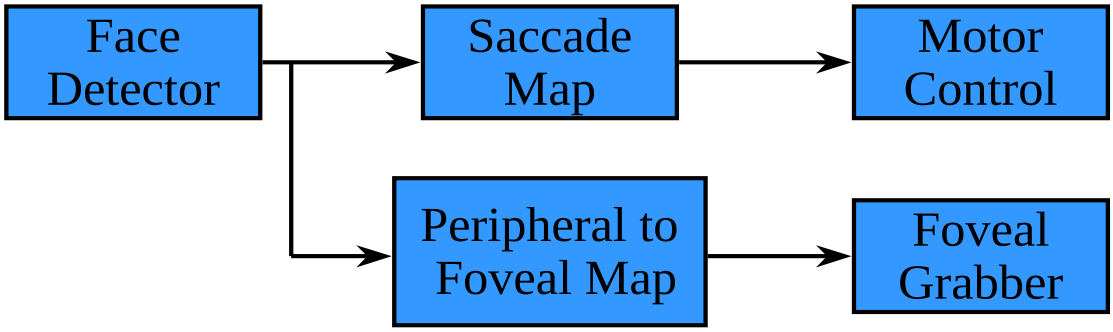

Fig. 5. Block diagram for finding eyes and faces. Once a target face has been located, the system must saccade to that location, verify that the face is still present, and then map the position of the eye from the face template onto a position in the foveal image.

are then mapped into foveal camera coordinates 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 (see Figure 5). This extracted image is then ready for further processing. The left image of Figure 6 shows the result of the face detection routines on a typical grayscale image before the saccade. The right image of Figure 6 shows the extracted subimage of the eye that was obtained after saccading to the target face. Additional examples of successful detections on a variety of faces can be seen in Figure 7. This method achieves good results in a dynamic real-world environment; in a total of 140 trials distributed between 7 subjects, the system extracted a foveal image that contained an eye on 131 trials (94% accuracy). Of the missed trials, two resulted from an incorrect face identification (a face was falsely detected in the background clutter), and seven resulted from either an inaccurate saccade or motion of the subject (Scassellati 1998*b*).

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. Work on extracting the location of the pupil within the eye and the position of the head on the body has begun, but is still in progress.

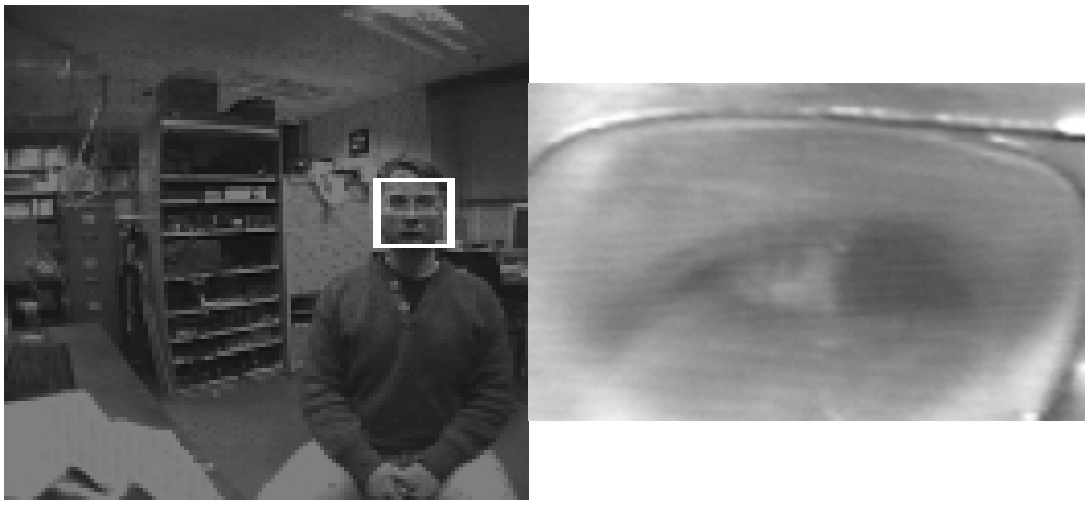

Fig. 6. A successfully detected face and eye. The 128x128 grayscale image was captured by the active vision system, and then processed by the pre-filtering and ratio template detection routines. One face was found within the peripheral image, shown at left. The right subimage was then extracted from the foveal image using a learned peripheral-to-foveal mapping.

3.2 Implementing Gaze Following

Once our system is capable of detecting eye contact, we require three additional subskills to achieve 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, and requires the skills noted in the preceding section. Extrapolation of the angle of gaze can be more difficult. 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 demonstrates a simpler strategy.

Butterworth (1991) has shown that at approximately 6 months, infants will begin to follow a caregiver's gaze to the correct side of the body, that is, the child can distinguish between the caretaker looking to the left and the caretaker looking to the right (see Figure 8). Over the next three months, their accuracy increases so that they can roughly determine the angle of gaze. At 9 months, the child will track from the caregiver's eyes along the angle of gaze until a salient object is encountered. 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, even at this stage, infants will only exhibit gaze

Fig. 7. Additional examples of successful face and eye detections. The system locates faces in the peripheral camera, saccades to that position, and then extracts the eye image from the foveal camera. The position of the eye is inexact, in part because the human subjects are not motionless.

following if the distal object is within their field of view. They will not turn to look behind them, even if the angle of gaze from the caretaker would warrant such an action. 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 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 objects that have salient color, intensity, or motion, we can mimic the ecological strategy. From an ecological mechanism, we can refine the algorithms for determining gaze and add mechanisms for determining vergence. A rough geometric strategy can then be implemented, and later refined through feedback from the caretaker. A representational strategy requires the ability to maintain information on salient objects that are outside of the field of view including information on their appearance, location, size, and salient properties. The implementation of this strategy requires us to make

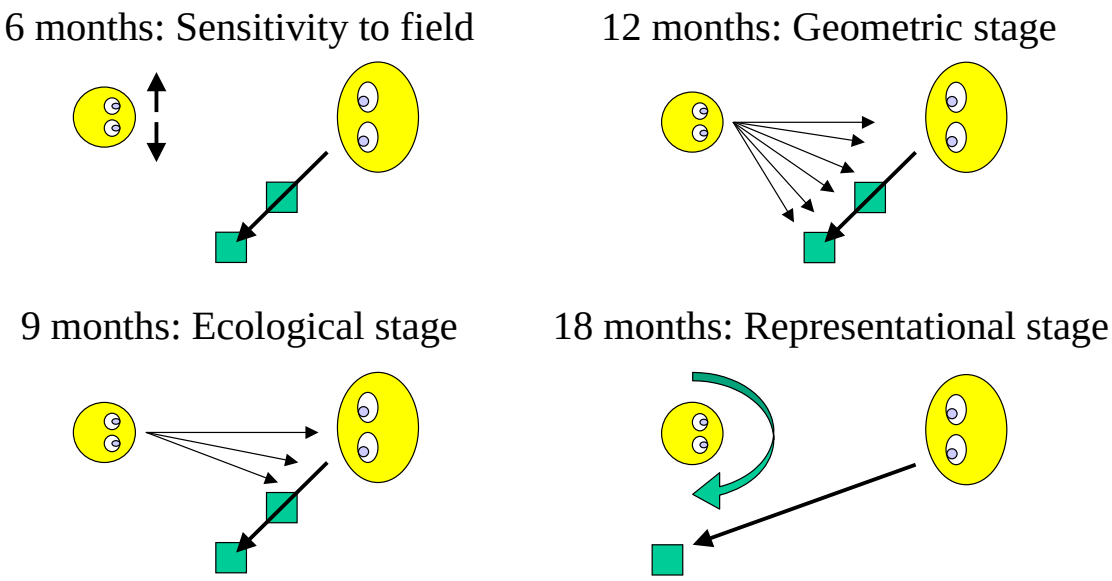

Fig. 8. Proposed developmental progression of gaze following adapted from Butterworth (1991). At 6 months, infants show sensitivity only to the side that the caretaker is gazing. At 9 months, infants show a particular strategy of scanning along the line of gaze for salient objects. By one year, the child can recognize the vergence of the caretaker's eyes to localize the distal target, but will not orient if that object is outside the field of view until 18 months of age.

assumptions about the important properties of objects that must be included in a representational structure, a topic beyond the scope of this chapter.

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 have utilized the foveation behavior obtained from the first step in order to train the arm where to reach (Marjanović et al. 1996). To reach to a visual target, the robot must learn the mapping from retinal image coordinates $\boldsymbol{x} = (x, y)$ to the head-centered gaze coordinates of the eye motors $\boldsymbol{e} = (pan, tilt)$ and then to the coordinates of the arm motors $\boldsymbol{\alpha} = (\alpha_0 ... \alpha_5)$ (see Figure 9). The saccade map $\boldsymbol{S} : \boldsymbol{x} \rightarrow \boldsymbol{e}$ 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 head-centered coordinates $\boldsymbol{e}$ to arm-centered coordinates $\boldsymbol{\alpha}$. To simplify

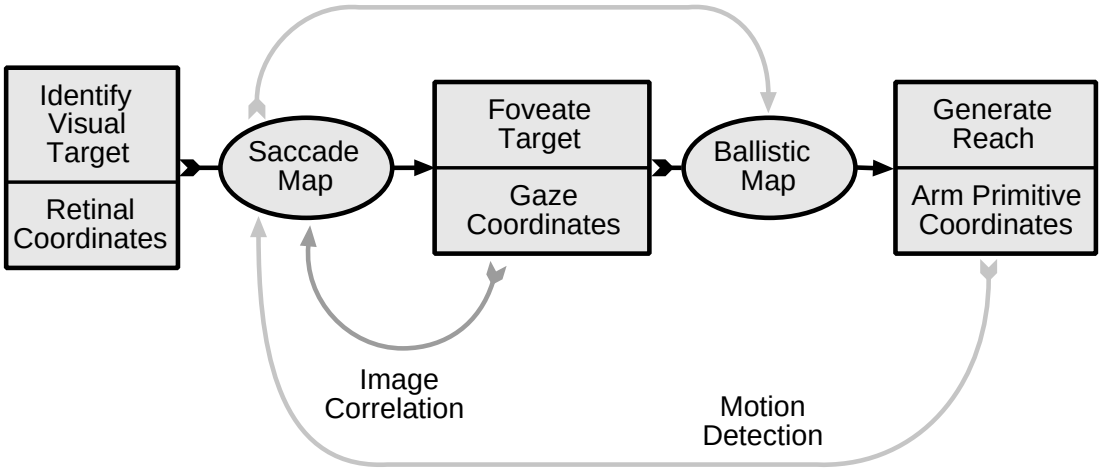

Fig. 9. Reaching to a visual target is the product of two subskills: foveating a target and generating a ballistic reach from that eye position. Image correlation can be used to train a saccade map which transforms retinal coordinates into gaze coordinates (eye positions). This saccade map can then be used in conjunction with motion detection to train a ballistic map which transforms gaze coordinates into a ballistic reach.

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 $\boldsymbol{B} : \boldsymbol{e} \rightarrow \boldsymbol{\alpha}$ 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. Once the saccade map has been trained, we can utilize that mapping to generate error signals for attempted reaches (see Figure 10). By tracking the moving arm, we can obtain its final position in image coordinates. The vector from the tip of the arm in the image to the center of the image is the visual error signal, which can be converted into an error in gaze coordinates using the saccade mapping. The gaze coordinates can then be used to train a forward and inverse model of the ballistic map using a distal supervised learning technique (Jordan & Rumelhart 1992). 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 reach using the ballistic map.
4. As the arm moves, use motion detection to locate the end of the arm.
5. Use the saccade map to convert the error signal from image coordinates into gaze positions, which can be used to train the ballistic map.
6. 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. Video clips of Cog reaching

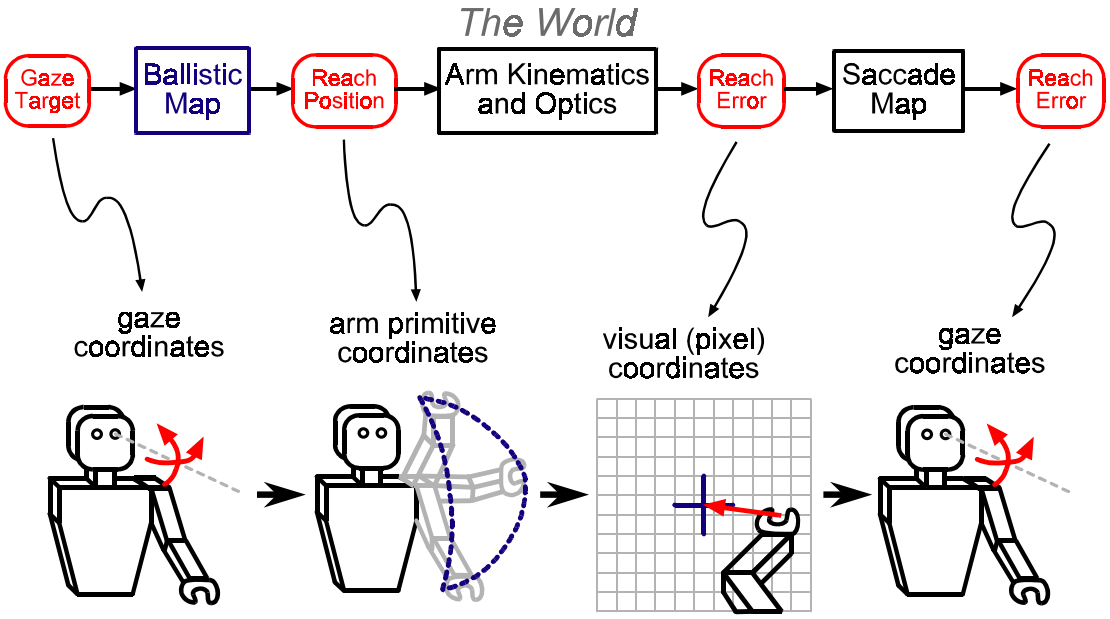

Fig. 10. Generation of error signals from a single reaching trial. Once a visual target is foveated, the gaze coordinates are transformed into a ballistic reach by the ballistic map. By observing the position of the moving hand, we can obtain a reaching error signal in image coordinates, which can be converted back into gaze coordinates using the saccade map.

to a visual target are available from `http://www.ai.mit.edu/projects/cog/`, and 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 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, that is, when the caretaker nods yes, the robot responds by nodding yes (see Figure 11). The face detection module produces a stream of face locations at 20Hz. An attentional marker is attached to the most salient face stimulus, and the location of that marker is tracked from frame to frame. If the position

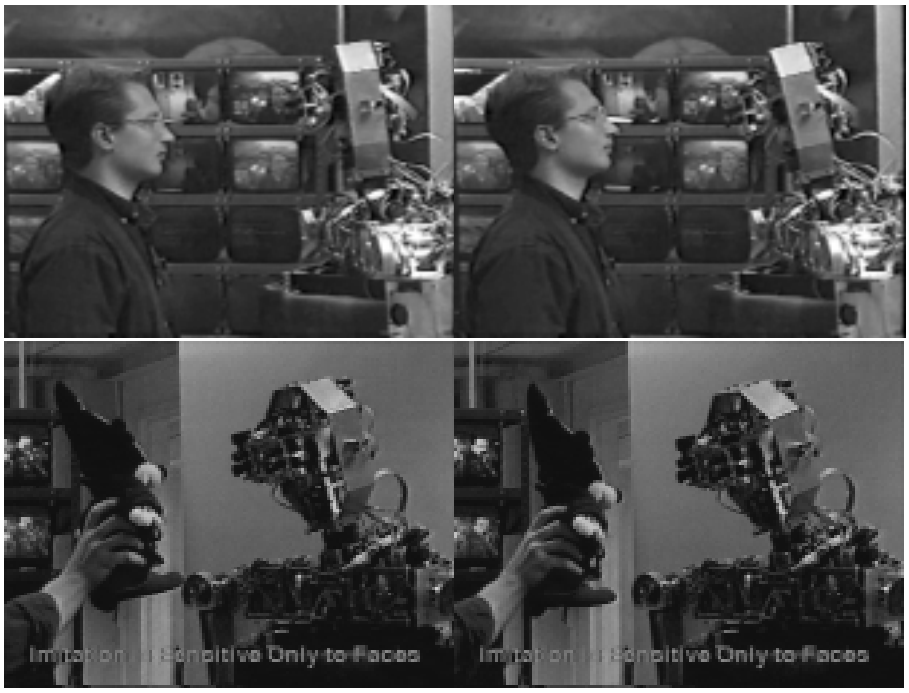

Fig. 11. Images captured from a videotape of the robot imitating head nods. The upper two images show the robot imitating head nods from a human caretaker. The output of the face detector is used to drive fixed yes/no nodding responses in the robot. The face detector also picks out the face from stuffed animals, and will also mimic their actions. The original video clips are available at `http://www.ai.mit.edu/projects/cog/`.

of the marker changes drastically, or if no face is determined to be salient, then the tracking routine resets and waits for a new face to be acquired. Otherwise, the position of the attentional marker over time represents the motion of the face stimulus. The motion of the attentional marker for a fixed-duration window is classified into one of three static classes: a *yes* class, a *no* class, and a *no-motion* class. Two metrics are used to classify the motion, the cumulative sum of the displacements between frames (the relative displacement over the time window) and the cumulative sum of the absolute values of the displacements (the total distance traveled by the marker). If the horizontal total trip distance exceeds a threshold (indicating some motion), and if the horizontal cumulative displacement is below a threshold (indicating that the motion was back and forth around a mean), and if the horizontal total distance exceeds the vertical total distance, then we classify the motion as part of the *no* class. Otherwise, if the vertical cumulative total trip distance exceeds a threshold (indicating some motion), and if the vertical cumulative displacement is below a threshold (indicating that the motion was up and down around a mean), then we classify the motion as part of the *yes* class. All other motion types default to the *no-motion* class. These simple classes then drive fixed-action patterns for moving the head and eyes in a yes or no nodding motion. 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, as well as further documention, are available from `http://www.ai.mit.edu/projects/cog/`.

4 Conclusion

Guided by evidence from developmental psychology, from disorders of social development such as autism, and from the evolutionary development of these skills, we have described a task-based decomposition of joint attention skills. Our implementation of this developmental progression is still in progress, but our initial results with finding faces and eyes, and with the imitation of simple head movements, suggest that this decomposition may be a useful mechanism for building social skills for human-like robots. If this implementation is successful, we can then begin to use the skills that our robot has acquired in order to test the developmental models that inspired our program. A robotic implementation will provide a new tool for investigating complex interactive models that has not been previously available.

5 Acknowledgements

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

The author wishes to thank the members of the Cog group for their contributions to this work: Rod Brooks, Cynthia Breazeal (Ferrell), Robert Irie, Charles Kemp, Matthew Marjanovic, and Matthew Williamson.

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.

Cohen, D. J. & Volkmar, F. R., eds (1997), *Handbook of Autism and Pervasive Developmental Disorders*, second edn, John Wiley & Sons, Inc.

Dautenhahn, K. (1994), Trying to Imitate — A Step Towards Releasing Robots from Social Isolation, *in* 'Proc. From Perception to Action Conference (Lausanne, Switzerland, Sept 7-9, 1994)', IEEE Computer Society Press, pp. 290–301.

Dautenhahn, K. (1997), 'I could be you — the phenomenological dimension of social understanding', *Cybernetics and Systems* **25**(8), 417–453.

Dennett, D. C. (1991), *Consciousness Explained*, Little, Brown, & Company.

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.

DSM (1994), 'Diagnostic and Statistical Manual of Mental Disorders', American Psychiatric Association, Washington DC.

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

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

Hayes, G. & Demiris, J. (1994), A Robot Controller Using Learning by Imitation, *in* A. Borkowski & J. L. Crowley, eds, 'Proc. 2nd International Symposium on Intelligent Robotic Systems', Grenoble, France: LIFTA-IMAG, pp. 198–204.

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

ICD (1993), 'The ICD-10 Classification of Mental and Behavioral Disorders: Diagnostic Criteria for Research', World Health Organization (WHO), Geneva.

Jordan, M. I. & Rumelhart, D. E. (1992), 'Forward Models: supervised learning with a distal teacher', *Cognitive Science* **16**, 307–354.

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.

Moore, C. & Dunham, P. J., eds (1995), *Joint Attention: Its Origins and Role in Development*, Erlbaum.

Povinelli, D. J. & Preuss, T. M. (1995), 'Theory of Mind: evolutionary history of a cognitive specialization', *Trends in Neuroscience*.

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.

Thelen, E. & Smith, L. (1994), *A Dynamic Systems Approach to the Development of Cognition and Action*, MIT Press, Cambridge, MA.

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

Wood, D., Bruner, J. S. & Ross, G. (1976), 'The role of tutoring in problem-solving', *Journal of Child Psychology and Psychiatry* **17**, 89–100.