Mechanisms of Shared Attention for a Humanoid Robot

Brian Scassellati
scaz@ai.mit.edu
MIT Artificial Intelligence Lab
545 Technology Square, Room NE43-835
Cambridge, MA 02139

Abstract

This paper outlines a proposal for constructing mechanisms of shared attention for a humanoid robot through a series of example tasks. Shared attention, the ability to selectively attend to objects that are mutually interesting, is vital for learning from another individual. The platform that we will use in designing this system is the upper-torso humanoid robot called Cog which is currently under construction at the MIT Artificial Intelligence Lab. We present a description of our approach, an outline of the work in progress, and a summary of the work completed so far.

Introduction

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 advent of embodied systems research, we are now capable of building systems that are robust enough, safe enough, and stable enough to allow machines to interact with humans in a learning environment.

One difficulty in enabling a machine (or a person) to learn from an instructor is ensuring that the student and the instructor are both attending to the same object of interest. When the instructor gives new information, the student must know what to apply that information towards. In other words, the student must know which parts of the scene are relevant to the lesson at hand and which parts are irrelevant. As we build more complicated embodied systems, with more sensors and more mobility, and with more general goals, this problem only becomes more complex. How can we sort out which sensory signals are important to the lesson at hand? How do we focus our attention on the important aspects of our sensations? Human students use a variety of social cues from the instructor for directing their attention; linguistic determiners (such as "this" or "that"), gestural cues (such as pointing or eye direction), and postural cues (such as proximity) can all direct attention to specific objects and resolve this problem. I will call the ability to selectively attend to the objects indicated by the instructor a mechanism of shared attention, that is, a method by which the instructor and the student both attend to the same object.

This paper outlines a proposal for the construction of mechanisms of shared (or joint) attention for a humanoid robot interacting with human instructors in a natural, unconstrained setting. The physical system that we will use for this project is an upper-torso humanoid robot called Cog, currently under construction at the MIT Artificial Intelligence Laboratory. Cog was designed to explore a wide variety of problems in artificial intelligence and cognitive science (Brooks & Stein 1994). To date our hardware systems include a ten degree-of-freedom upper-torso robot, two multiprocessor MIMD computers, a binocular, foveated video capture/display system, a six degree-of-freedom series-elastic actuated arm, and a host of programming language and support tools (Brooks 1996; Brooks *et al.* 1996). Additional information on the project background can be found in (Brooks & Stein 1994; Marjanović, Scassellati, & Williamson 1996; Ferrell 1996; Williamson 1996; Irie 1995; Marjanovic 1995; Matsuoka 1995; Pratt & Williamson 1995; Scassellati 1995).

Examples of Shared Attention Mechanisms

Before proceeding further, let us consider two scenarios of shared attention that Cog should be able to perform: the ability to achieve shared attention by reacting to the instructor and the ability to request shared attention from the instructor. Because Cog currently lacks auditory and speech production capabilities, we will limit ourselves to non-linguistic mechanisms of shared attention.

Cog should be able to share attention with an instructor by responding to many different cues. One common cue for shared attention is direction of gaze.

For example, a person who is looking at you is more likely to be addressing you than a person looking elsewhere. Also, a person who is looking above you may have seen something interesting that you might want to look at. In this situation, Cog must extract the direction of gaze of the instructor to obtain the relevant attentional cue. Another variant of this situation is to achieve shared attention from a pointing gesture. Cog would need to determine what the instructor is pointing at in order to achieve shared attention. In both cases, Cog needs the capabilities to recognize and respond to the natural social attentional cues that the instructor displays.

Cog should also be able to request shared attention from the instructor. By pointing at an interesting object, or looking back and forth between the instructor and the object, Cog can request that the instructor attend to an object. To complete the transaction, Cog must also be able to recognize when the instructor's attention has reached the same object; thus, this second scenario may be built upon the first.

Design Tenets for Implementing Shared Attention

One of the central design tenets of the Cog project is to learn about the world by interacting with the environment. There should not be a large corpus of "hard-coded" rules and regulations that are input to the robot; the robot should be capable of learning experientially by interacting with the world. We will adopt this as our first design tenet in implementing shared attention. Our mechanisms of shared attention should consist of behaviors that are acquired through successful interaction with the instructor, and contain few invariants.

The second tenet of our proposal will be to use the same social cues of shared attention that humans use; we will not construct some artificial grammar that must be learned in order to interact with the robot. The use of natural cues has many advantages: Our instructors will need no special training, our choice of mechanisms will not limit the types of problems we can learn, and we can use knowledge about human interaction as guidelines for constructing our system. Cog was designed to be anthropomorphic in its appearance and its capabilities so that it could interact with people using these common social cues. In the next section, we begin to explore the structure of shared attention in people as a starting point for our implementation.

Shared Attention in Biological Systems

The mechanisms for shared attention in humans are not a single monolithic system. Evidence from childhood development shows that not all mechanisms for shared attention are present from birth. There are also developmental disorders that limit and fracture this system. 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 developmental behavior. It is possible that Cog will not only be an interesting computational and mechanical exercise, but also a foil to test theories of developmental progress in humans.

Normal Human Development

Studying the development of shared attention mechanisms in children is a difficult proposition. Because shared attention begins developing in the first year, it is difficult to distinguish between failures that are a result of lack of motor control, failures due to lack of underlying cognitive abilities, and failures of executive or inhibitory control (Diamond 1990). However, it does seem clear that not all parts of the mechanisms of shared attention are innate. The social cues of gesture and posture differ slightly between cultures, so some post-natal learning is necessary.

Normal children do not begin to show some of the aspects of shared attention until the second or third year of life (Hobson 1993). Children younger than nine months do not show reliable gaze monitoring. Not until 12 months do infants begin to show proto-declarative pointing (pointing to indicate an object that they want), and it is not until 18 months that they can track another person's line of sight outside their own visual field. It is not until four years of age that children can reliably predict what another individual can or cannot see. This gradual development may be the precursor to more sophisticated concepts such as a "theory of mind," that is, the ability to reflect on one's own mental states and the mental states of others (Baron-Cohen 1994). The part that shared attention mechanisms play in this development in normal children is not well understood, but the absence of these mechanisms is devastating. (See the section on abnormal development below.)

Evolutionary Evidence: Primates

Studying the development of mechanisms of shared attention in normal children provides one way of decomposing this system. Another method is to take an evolutionary perspective and look at the abilities of shared attention that are present in other primates. The evolutionary advantages of shared attention to a social creature are enormous; shared attention allows the individual to cooperate with others, signal the locations of food or danger, or to predict the behavior of others.

Other primates show varying degrees of sophistication in their mechanisms of shared attention (Povinelli & Preuss 1995). All primates recognize and exhibit proto-declarative pointing. Only chimpanzees and the great-apes will exhibit gaze following, that is, they will shift their posture and gaze to follow the line of sight of another. However, chimps fail to recognize the implications that seeing is knowing. For example, if the chimp is forced to choose between taking the advice of a trainer who saw where food was being hidden and taking the advice of a blindfolded trainer, chimps behave at chance levels in initial trials. The chimps eventually learn this "trick" and choose the trainer who saw the food being hidden, but do not generalize the skill. [1] Despite differences in interpretation of social cues, the evidence from primates indicates that the mechanisms of shared attention can be decomposed into different abilities.

Abnormal Human Development

In addition to studying normal development and evolutionary development of the mechanisms of shared attention, there are also interesting cases in abnormal child development that provide information on decomposing this system. For example, children with congenital blindness lack all of the visual mechanisms, but still learn to develop normal mechanisms for shared attention. Without intensive training, they learn the normal social signals carried by pitch, inflection, and other subtleties of speech. There is also evidence that they understand the visual mechanisms of shared attention even if they are unable to experience them. Blind children who are asked to show an object to an adult will hold out that object for inspection; if they are asked to hide an object, they will properly place the object out of the line of sight of the observer (Frith 1990).

Perhaps the most interesting case for shared attention mechanisms comes from the study of autism. Autistic children seem to lack some of the mechanisms of shared attention. For example, although autistics have no difficulty in detecting eye direction, they fail to show normal forms of gaze following. Autistics have no difficulty in identifying that a person is looking at them, or at a specific object, but they fail to associate that cue in terms of belief or desires, or to engage in shared attention through gaze following or monitoring. This is clearly demonstrated by a forced choice task often called the "Smarties" task (named after a popular British candy). Children are shown a picture of a person surrounded by four types of candy. When normal children are asked which candy the person wants, they will answer with whatever candy the person is looking at. Autistic children answer this question randomly, or with their own candy preference, and yet have no difficulty in answering the question: "Which candy is X looking at?" While the deficits of autism certainly cover many other areas of perceptual and cognitive capabilities, 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 *et al.* 1995).

[1] It should also be said that the social signals of shared attention are interpreted differently in non-human primates. For example, to non-human primates, looking directly into the eyes of another is a signal of social aggression.

Implementing a Modular Mechanism of Shared Attention

Baron-Cohen's Model

One of the most comprehensive models of shared attention comes from Simon Baron-Cohen's work with autism (Baron-Cohen 1995). Baron-Cohen describes a three-tiered model containing four modules which combine to produce behavior that matches the shared-attention capabilities of normal four-year-olds. In the first tier are two independent modules, the intentionality detector (ID) and the eye direction detector (EDD). ID is a multi-modal module that recognizes stimuli with self propulsion through visual, auditory, and tactile stimuli, and produces goal/desire relationships with objects [*he wants X*]. EDD is a visual module that detects eye-like stimuli and produces dyadic relationships about visibility [*he sees X*]. The results from these two modules are passed to the what Baron-Cohen calls the "shared attention module" (SAM). SAM links the information from ID and EDD into a triadic representation ([Mommy sees [he sees X]] and then [Mommy sees [he wants toy]]). Finally, this triadic representation is given to a catch-all module called the theory-of-mind-module (TOMM). TOMM is capable of representing the full range of mental and attentional concepts.

What makes Baron-Cohen's model interesting is that the four modules are separable based on the developmental evidence discussed above. Both EDD and ID are present in normal children by 9 months. SAM develops between 9 and 18 months, while TOMM may not develop fully until 36 to 48 months. Also, while EDD is obviously absent in blind children, the output from ID is sufficient to drive the rest of the system. Baron-Cohen also proposes that both EDD and ID are present in chimps, as well as some aspects of SAM, but not TOMM. For autistics, it seems that ID and EDD are operational, but that the functions of SAM (and thus TOMM) are impaired. The simple breakdown of the shared attention mechanisms described by this model matches the biological evidence.

Proposed Implementation

What Baron-Cohen's model does not provide is a task-level decomposition of what skills are necessary to provide the functionality of his modules. The description that follows is an account of a few subtasks that will build upon each other to construct a mechanism of shared attention for Cog. While this account will certainly not enable Cog to perform the full range of theory-of-mind tasks that Baron-Cohen describes at the upper-most tier in his model, nor will it conform strictly to the functional decomposition of his model, we should be able to build an embodied system that allows for more general learning algorithms to take advantage of the opportunities presented by a helpful instructor.

The first step in producing mechanisms of shared attention will be gaze monitoring. Just as human infants have an innate preference for looking at human faces and eyes, Cog should show a hard-wired preference to look at human faces and eyes. This first scenario requires a great deal of perceptual capability, motor control, and sensorimotor integration. Simply maintaining eye contact requires the ability to detect eyes and/or faces, to maintain eye-motor fixation, and to smoothly track visual objects.

The second step will be to engage in shared attention by interpolation of gaze. In all of the great apes, when an instructor moves its gaze to a new location, the student will move its gaze to that same location. In addition to the perceptual and motor skills required from our first step, this added functionality requires detecting direction of gaze and interpolating that gaze angle in the scene (perhaps with knowledge of depth). How can we move from gaze monitoring to shared attention through gaze? One possibility is to use a preference for motion to learn the predictive relationship of shared attention through gaze. Gaze angle of the instructor is likely to be an excellent predictor of moving objects, or objects that will later be presented to the student. An alternative possibility is to use positive reinforcement from the instructor. By attending to objects that the instructor indicates, the pupil can be rewarded with more attention, with the object itself, or through the fulfillment of some other internal desire. However, placing the burden of learning this relationship on the individual may be unnecessary; these relationships are likely to be instinctual. In all social animals, the instinct to use the observations, fear responses, and direction of gaze of other pack members is very basic. It is possible that this piece should be an innate part to our system. With the completion of this step, we have obtained most of the properties of Baron-Cohen's EDD and ID modules for a limited domain (gaze).

The third step in our account will be to engage in shared attention through pointing. Just as gaze direction can indicate a request for shared attention, a pointing gesture is a request for the student to attend to a distal object. Adding pointing as a secondary source of information requires visual posture identification and classification in addition to the requirements from the second step. Learning that pointing is a request for shared attention is remarkably similar to learning that gaze is a request for shared attention; this may come as a result of positive reinforcement or regularities in its predictive capability, but can also be attributed to evolved instincts. The addition of pointing adds a secondary modality to the repertoire of EDD and ID, but does not substantially alter the functional structure.

The final step in accounting for our two example scenarios is to enable Cog to *request* shared attention through pointing. This step adds more complex computational and motor control requirements to our system. For Cog, pointing to an object requires coordination between at least 12 degrees of freedom.[2] Furthermore, the system will require a more detailed visual targeting system to identify, saccade to, and maintain fixation of visual objects.[3] How to construct requests for shared attention is a challenging problem. The possibility that I am currently exploring is learning to request shared attention through "deep" imitation and positive reinforcement. By "deep" imitation, I mean to signify that Cog will not simply mimic the actions of the instructor. Instead, Cog will imitate the functional gesture that the instructor performs. In simple mimicry, if the instructor were to point to his head, then Cog should also point to its own head. In deep imitation, Cog would imitate not the exact posture, but rather the intended functional goal of that gesture; instead of pointing to its own head, Cog would point to the instructor's head. By using deep imitation and positive reinforcement, it is possible to learn to request shared attention through observing (and imitating) shared attention requests. For Cog to be capable of deep imitation, it is necessary to include the geometric interpolations that the previous stages have developed as well as a functional mapping of the instructor's posture onto Cog's body.

Once these four steps have been implemented, Cog should be capable of learning from people in a natural, unconstrained manner. These mechanisms of shared attention will allow our embodied system to continue to learn from its environment by learning from the people in that environment. Just as a child learns social skills and conventions through interactions with its parents, Cog will learn to interact with people using natural, social communication.

Current Work

The implementation of the perceptual, motor, and sensorimotor integration tasks necessary to build mechanisms of shared attention for Cog are still in progress. We currently have operational systems that provide basic eye-motor control, smooth tracking, visual motion detection, and learning algorithms for determining the sensorimotor mapping to saccade to a visual target.

We have also implemented a self-taught visually-guided pointing algorithm that enables Cog to learn

[2] There are six degrees of freedom in the arm, three in the neck, and three in the eyes.

[3] For the time, we will not consider how objects are selected as worthy of shared attention. This can be the product of other basic drives or higher-level cognitive modules.

to manipulate 12 degrees of freedom to point to a visual target. The algorithm uses the visual motion of the arm to learn a mapping between the eye, neck, and arm postures and the visual scene. More information on this task can be found in (Marjanović, Scassellati, & Williamson 1996).

Conclusion

In this brief presentation, I have outlined a proposal for constructing mechanisms of shared attention for a humanoid robot. These mechanisms will be a vital aspect of enabling our robot to learn from an instructor using simple, natural social cues.

The implementation of these mechanisms will also be an interesting combination of machine learning, machine vision, and robotics. The perceptual, motor, and integrative tasks that we will require from Cog will be useful not only to this project, but to other tasks as well. The implementation of shared attention may also provide some insights to the feasibility of the models of normal and abnormal human development that Baron-Cohen has proposed.

Acknowledgments

The author wishes to thank the members of the Cog group (past and present) for their continual support: Rod Brooks, Cynthia Ferrell, Robert Irie, Matt Marjanovic, Yoky Matsuoka, Lynn Stein, and Matt Williamson.

This material is based upon work supported by a National Defense Science and Engineering Graduate Fellowship awarded to the author. Additional support for the Cog project is provided by the National Science Foundation under National Science Foundation Young Investigator Award Grant No. IRI–9357761 to Professor Lynn Andrea Stein. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Baron-Cohen, S. 1994. Current psychology. *Cognition.*

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

Brooks, R., and Stein, L. A. 1994. Building brains for bodies. *Autonomous Robots* 1:1:7–25.

Brooks, R.; Bryson, J.; Marjanovic, M.; Stein, L. A.; ; and Wessler, M. 1996. Humanoid software. Technical report, MIT Artificial Intelligence Lab Internal Document.

Brooks, R. 1996. L. Technical report, IS Robotics Internal Document.

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

Ferrell, C. 1996. Orientation behavior using registered topographic maps. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB-96).* Society of Adaptive Behavior.

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

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

Irie, R. 1995. Robust sound localization: An application of an auditory perception system for a humanoid robot. Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Karmiloff-Smith, A.; Klima, E.; Bellugi, U.; Grant, J.; and 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. J.; Scassellati, B.; and Williamson, M. M. 1996. Self-taught visually-guided pointing for a humanoid robot. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB-96).* Society of Adaptive Behavior.

Marjanovic, M. 1995. Learning functional maps between sensorimotor systems on a humanoid robot. Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Matsuoka, Y. 1995. Embodiment and manipulation learning process for a humanoid hand. Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Povinelli, D. J., and Preuss, T. M. 1995. Theory of mind: evolutionary history of a cognitive specialization. *Trends in Neuroscience.*

Pratt, G. A., and Williamson, M. M. 1995. Series elastic actuators. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-95)*, volume 1, 399–406.

Scassellati, B. 1995. High level perceptual contours from a variety of low level features. Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Williamson, M. M. 1996. Postural primitives: interactive behavior for a humanoid robot arm. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB-96).* Society of Adaptive Behavior.