Foundations for a Theory of Mind for a Humanoid Robot

by

Brian Michael Scassellati

S.B., Electrical Engineering and Computer Science
S.B., Brain and Cognitive Sciences
M.Eng., Computer Science and Electrical Engineering
Massachusetts Institute of Technology, 1995

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2001

© Massachusetts Institute of Technology 2001. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
May 6, 2001

Certified by . . . .
Rodney Brooks
Fujitsu Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Foundations for a Theory of Mind for a Humanoid Robot

by
Brian Michael Scassellati

Submitted to the Department of Electrical Engineering and Computer Science
on May 6, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Human social dynamics rely upon the ability to correctly attribute beliefs, goals, and percepts to other people. The set of abilities that allow an individual to infer these hidden mental states based on observed actions and behavior has been called a "theory of mind" (Premack & Woodruff, 1978). Existing models of theory of mind have sought to identify a developmental progression of social skills that serve as the basis for more complex cognitive abilities. These skills include detecting eye contact, identifying self-propelled stimuli, and attributing intent to moving objects.

If we are to build machines that interact naturally with people, our machines must both interpret the behavior of others according to these social rules and display the social cues that will allow people to naturally interpret the machine's behavior.

Drawing from the models of Baron-Cohen (1995) and Leslie (1994), a novel architecture called *embodied theory of mind* was developed to link high-level cognitive skills to the low-level perceptual abilities of a humanoid robot. The implemented system determines visual saliency based on inherent object attributes, high-level task constraints, and the attentional states of others. Objects of interest are tracked in real-time to produce motion trajectories which are analyzed by a set of naive physical laws designed to discriminate animate from inanimate movement. Animate objects can be the source of attentional states (detected by finding faces and head orientation) as well as intentional states (determined by motion trajectories between objects). Individual components are evaluated by comparisons to human performance on similar tasks, and the complete system is evaluated in the context of a basic social learning mechanism that allows the robot to mimic observed movements.

Thesis Supervisor: Rodney Brooks
Title: Fujitsu Professor of Computer Science and Engineering

Acknowledgments

Cog has been an enormous engineering and scientific project which has benefited from the efforts and ideas of many, many people.

This project simply could never have happened without Rod Brooks. Rod has created an environment at the MIT AI lab where people can work on ideas and projects that would be considered far too ambitious, too outlandish, or too controversial somewhere else. He has given me the freedom to tackle big questions and the support to recover from big mistakes. Thank you, Rod.

I am also indebted to the other members of my committee. Eric Grimson has been a source of both practical advice and endless inspiration since my first registration day as a graduate student. Leslie Kaelbling has held me to a high standard, even when faced with rapidly approaching deadlines and many other commitments. This document would not have been complete without her input.

Cynthia Breazeal has been my longtime collaborator and a kindred spirit in the pursuit of social machines. She is the guiding hand behind the robot Kismet, and was part of the original Cog team. I have learned much from her.

Thanks also to the members of the Cog team, past and present. Matt Marjanovic has been there since the beginning, constantly pushing to put together a stable research platform. I promise not to make you switch processors ever again. Bryan Adams has been invaluable in making Cog's arms a usable resource, and in keeping me from going stark raving mad. Aaron Edsinger has become our new mechanical guru, building new hands and reworking the robot's torso in the last year. The four of us have built extensively on the many previous iterations of hardware and software design. Matt Williamson, Mike Binnard, Cynthia Breazeal, and Robert Irie were integral parts of this process.

The other members of the Humanoid Robotics Group at the MIT Artificial Intelligence Lab have been a daily source of inspiration and support. Paul Fitzpatrick, Lijin Aryananda, and Paulina Varchavskaia have helped to make Kismet one of the most engaging machines ever constructed. Aaron Edsinger, Giorgio Metta, and Una-May O'Reilly have been pushing the boundaries of both our visual processing software and our understanding of how people interact with Lazlo. Thanks also to Jessica Banks, Charlie Kemp, and Juan Velasquez for working the late nights with me. It was much more fun with all of you there. And finally, thanks to Eduardo Torres, Artur Arsenio, Naoki Sadakuni, Chris Morse, Holly Yanco, James McLurkin, and everyone else who I've had the pleasure of working with in the past few years.

There have also been many visitors to our group at MIT in the past seven years. They have all contributed to the atmosphere of excitement that has surrounded these projects throughout my time here. Special thanks to Giulio Sandini, Hideki Kozima, Yasuo Kuniyoshi, and Manuela Veloso in particular for sharing their perspectives and enriching my work.

And then there are the others who have contributed in subtle, but important ways. Maja Mataric and Ian Horswill have been mentors from afar, always encouraging me to push one step further. Dan Dennett has been a steady hand in guiding my thoughts and in exploring the implications of the work that I do. Kerstin Dautenhahn was

the first to really show me that there was interest in what I was doing outside the robotics community. Sherry Turkle and Jen Audley have renewed my excitement and wonder at how children become invested in these technologies. Roz Picard and Bruce Blumberg have offered encouragement and wisdom. Pawan Sinha and Ingemar Cox kindly allowed use of their software for face detection and motion correspondence, respectively. Thanks to everyone.

I want to thank my friends and family for helping to encourage me through these many years.

Most importantly, none of this would be possible without the love and support of my wife, Kristi Hayes. She has the skill and the talent to do so many things that are difficult or impossible for me. She also still manages to help me with the things that I do. I am constantly in awe of her.

I am grateful for the commitment that funding agencies have placed in maintaining a program that was always a mixture of applied engineering and basic science. Portions of this research were funded by DARPA/ITO under contract number DABT 63-99-1-0012, “Natural tasking of robots based on human interaction cues,” and in part by an ONR/ARPA Vision MURI Grant (No. N00014-95-1-0600). I also received support from a fellowship from the Bose Foundation and by a National Defense Science and Engineering graduate fellowship. I hope that each of these organizations is pleased with the result.

Contents

List of Figures

Chapter 1

Introduction

> *Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?* – Turing (1950, p. 456)

1.1 A Grand Challenge: Social Robots

Many researchers have written about the problems of building autonomous robots that can perform useful tasks in environments that are hazardous to human beings. Whether searching for victims trapped in a destroyed building, inspecting nuclear waste sites, detecting land mines, or exploring the surface of another planet, these robots face environments that are extremely complex, unstructured, and hostile. For these systems, the designer has no control over the environment and cannot rely upon simplifying assumptions (such as static backgrounds or known physical environments) that make other robotics research practical. Programming systems to deal with all of the possible contingencies for such a complex and unstructured environment is an enormous task – programming robots to perform even simple tasks in structured environments is generally a large software engineering project. Rather than attempting to explicitly program the robot with responses for a large number of events, an alternative approach is to provide the robot with the capability to autonomously acquire the information that is required to make these behavioral decisions – in short, to learn. However, the same complexities of the environment that lead to the consideration of learning machines produce situations that most learning techniques are very poorly equipped to handle. These hostile environments present the robot with a wide variety of possible actions and yet only a few of these actions are generally beneficial. A single failure in these hostile environments can have very serious consequences, resulting in the destruction of the robot or injury to human beings.

The environments faced by these robots are very different from the nurturing environment in which human children are (typically) raised. Parents spend an enormous amount of time, energy, and resources on constructing an environment that is both safe and supportive for the child. Parents ensure that the environment contains few physical hazards for the child while still allowing the child enough freedom to explore different actions and to learn to recognize dangerous situations. Both the environment and the tasks that the child must perform are scaled to the level of ability of the child; adults do not attempt to teach algebra to a two-year-old child. The child also receives almost continuous feedback on the actions that it performs through the words, gestures, and expressions of the adult. In many ways, the child's environment is very well designed to support learning. The fundamental difference between these two environments is the presence of a helpful, knowledgeable caregiver. It is the parent that changes the otherwise hostile or indifferent environment into one in which the child can flourish. The child is able to learn effectively because of the structure provided by the parent.

In many ways, the idea of a machine that can learn from its own interactions with the world has been one of the driving forces behind artificial intelligence research since its inception (Turing, 1950). The most powerful form of this grand challenge is an android, a robot shaped like a human, that could master new skills and abilities by interacting with a person in the same way that you or I might attempt to learn a new skill from another person. This mythical machine could be taught a wide variety of skills with no more effort from the instructors a human student would require. This machine would be able to exploit the knowledge and assistance of other people to carry out specified tasks, would recognize and respond to the appropriate human social cues, and would use the natural social interfaces that people use with each other. A person would need no specialized training in order to instruct the machine, although some individuals would naturally be better instructors than others. To be an effective student, the robot would need many different skills. It would need a rich set of perceptual abilities for perceiving complex social cues and for guiding its behavior toward objects in the environment. A robust collection of behavioral responses for manipulating objects in the world, for performing tasks in the environment, and for safely engaging in cooperative tasks with people would be essential, as would a wide range of cognitive skills for planning, selecting appropriate actions, and for learning from past experiences. The robot would also need to be capable of producing the social cues that the human instructor required either to maintain the interaction dynamics or to evaluate the understanding and progress of the student.

Many different research fields have important contributions to make toward this grand challenge. Even when concentrating on the abilities required for the social learning aspects of the system (and excluding all of the abilities required for actually performing interesting tasks and for maintaining the survival of the system), there are still a wide range of disciplines that contribute to this ability. Research from machine vision, auditory analysis, and signal processing would contribute perceptual abilities for locating the instructor, recognizing the actions being performed, identifying objects, and understanding speech. Existing research in artificial intelligence that focuses on behavior selection and action planning would interact with motion control research on constructing safe, flexible, and robust behavioral responses with low latency. All of these systems would require real-time responses; research in embedded systems on parallel algorithms and real-time control would be applicable. Research on human-machine interfaces would guide the generation of appropriate social responses and the generation of social scripts. Of course, machine learning techniques for building sequences of actions and for using feedback cues to improve performance would be central to this endeavor. Even with this multitude of fields of research contributing to this challenge, the set of skills that can be assembled from existing research does not completely address the problems of social learning. There are many additional problems that are unique to this challenge that are not currently being studied. For example, how does the robot recognize when the social context is appropriate for learning? How does the robot recognize which actions it should be learning? Can the robot recognize and respond to the goal of an action rather than the surface behavior?

The grand challenge of building machines that can learn naturally from their interactions with people raises many difficult questions, but also offers the hope of overcoming the scaling problem.

1.2 A Specific Challenge: Theory of Mind

One area which has not received a great deal of attention from the computer science community, but which has been studied extensively in other fields, often goes by the name "theory of mind" (Premack & Woodruff, 1978). As we observe the behavior of other people, we naturally attribute to them beliefs, goals, percepts, and other mental states that we cannot directly observe. In the terms of computer science, theory of mind is the attempt to represent the hidden state maintained by another agent based upon the observable behavior of that agent. This set of abilities is also sometimes known as the ability to "mentalize" (Frith & Frith, 1999) or the ability to "mindread" (Baron-Cohen, 1995). Human social dynamics are critically dependent on the ability to correctly attribute beliefs, goals, and percepts to other people. A theory of mind allows us to understand the actions and expressions of others within an intentional or goal-directed framework (what Dennett (1987) has called the intentional stance). The recognition that other individuals have knowledge, perceptions, and intentions that differ from our own is a critical step in a child's development and is believed to be instrumental in self-recognition, in providing a perceptual grounding during language learning, and possibly in the development of imaginative and creative play (Byrne & Whiten, 1988). These abilities are also central to what defines human interactions. Normal social interactions depend upon the recognition of other points of view, the understanding of other mental states, and the recognition of complex non-verbal signals of attention and emotional state.

A robotic system that possessed a theory of mind would allow for social interactions that have previously not been possible between man and machine. The robot would be capable of learning from an observer using normal social signals in the same way that human infants learn; no specialized training of the observer would be necessary. The robot would also be capable of expressing its internal state (emotions, desires, goals, etc.) through social interactions without relying upon an artificial vocabulary. Further, a robot that could recognize the goals and desires of others would allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly. For example, Wood et al. (1976) have proposed that theory of mind is critical to learning language. By attending to the attitude and intent of the parent, the child first learns which situations are potential teaching episodes (marked by long extents of alternating eye contact and distal attention). The child then learns to associate specific utterances with the parent's object of attention. By attending to these social cues, the infant is able to determine which object is under consideration and can apply the new utterance selectively to that object.

Researchers from many fields have attempted to delineate the skills that serve as a foundation for a theory of mind. Developmental psychologists study how normal children acquire skills such as making eye contact or pointing to an object of interest. Ethologists consider the presence or absence of these abilities in other species. Researchers of developmental disorders study how these social abilities are either absent or impaired in certain conditions such as autism and Asperger's syndrome. These researchers have focused on behaviors that allow for the recognition of important social cues (such as gaze direction), behaviors that generate appropriate social cues (such as pointing to a desired object), and cognitive skills that attribute high-level concepts of animacy and intent. This endeavor is really an attempt to link what have classically been considered to be mysterious, high-level cognitive skills to actual behavioral triggers. A study of the foundations of a theory of mind is really an attempt to link low-level perceptual capabilities with a high-level cognitive model.

One of the critical aspects of any system that hopes to approach the grand challenge of social machines will be a theory of mind. Theory of mind skills have been studied in many disciplines as a way of bridging between high-level cognitive phenomena and low-level perceptual systems. Constructing the foundational skills for a theory of mind will investigate the link between these two realms.

1.3 Overview

The work presented in this thesis is an attempt to construct an embodied system capable of performing many of these foundational skills for a theory of mind. The implementation will be based on models of the development of theory of mind skills which account for behaviors observed in normal children, in autistic individuals, and in other primates. The goal of this implementation is two-fold: to provide an engineering implementation that can support social learning mechanisms by responding appropriately to basic social cues, and to provide an existence proof for a novel model of how these foundational skills interact. It should be made clear at this point that the work presented here is *not* being proposed as an explicit model of how theory of mind develops in humans. Although the work presented here is based extensively on models of human performance, the success of this model in presenting similar behavior on the robot does not imply that similar behavior observed in humans results from the same underlying structure; just because the model works on robots does not mean that people necessarily work the same way. However, the model will provide a proof of concept that certain aspects of popular human models may not be necessary to generate the observed behaviors. Chapter 12 will return to the questions of what implications can be drawn from this model in particular and how a robotic implementation might be used as a predictive tool for cognitive models of human behavior.

The implementation presented here focuses on three foundational skills for a theory of mind:

- Attribution of Animacy : The ability to distinguish between animate and inanimate objects on the basis of the spatial and temporal properties of their movement.

- Joint Attention : The ability to direct attention to the same object to which someone else is attending.

- Attribution of Intent : The ability to describe the movement of pairs of objects in terms of simple intentional states such as desire or fear.

These three abilities are implemented to operate at real-time interaction rates on a humanoid robot. To further demonstrate the effects of these foundational skills on social learning, these foundational skills were integrated with a system that allows the robot to mimic the movement of agents or objects in the environment.

The outline of the remainder of this document is as follows:

- *Chapter 2 : Methodology*

  We begin with a discussion of the general methodology that has been employed in building social robots using models of human development. Certain assumptions about the nature of human intelligence that are commonly found in classical artificial intelligence research are discarded and an alternative set of qualities are emphasized including physical embodiment, social interaction, integration, and developmental progression.

- *Chapter 3* : Developmental Models of Theory of Mind

  With this methodological foundation in place, we review data on how normal children develop theory-of-mind skills, how these same skills are deficient in individuals with certain developmental disorders (such as autism), and how some of these skills are observed in other animals. Two popular models of the development of theory of mind skills, one from Leslie (1994) and one from Baron-Cohen (1995), are reviewed and a novel hybrid model called *embodied theory of mind* is introduced.

- *Chapter 4* : Robot Platforms

  Three humanoid robots were constructed in part to support the implementation of the embodied theory of mind model. The capabilities of these robots that are relevant to social interaction are discussed in this chapter.

- *Chapter 5* : Matching Human Movements

  Once the physical structure of these robots has been described, we turn our attention to the kinds of behaviors that have been implemented on these robots to support social interaction. Human-like eye movements as well as arm movements such as pointing gestures were implemented to allow the robot to have natural interactions with people.

- *Chapter 6* : Visual Attention

  This chapter begins to consider how the robot's perceptual system has been structured to support social interaction. The first problem that the perceptual system must solve is to locate regions of interest that are worthy of further processing. The attention system selects regions based on inherent object properties (such as bright color or motion), high-level motivational goals, or the attentional states of others within the field of view.

- *Chapter 7* : The Theory of Body Module

  The attention system produces a set of interesting points for each object frame, which are then linked together using a motion correspondence algorithm to produce movement trajectories. The theory of body module applies a simple set of naive rules of physics to these movement trajectories in order to identify objects that are self-propelled. Self-propelled objects are considered to be animate, which will be treated as a special class in chapters 9, 10, and 11.

- *Chapter 8* : Detecting Faces and Head Pose

  One final primitive perceptual process will be required. The robot will need to find human faces in the visual scene and to extract the orientation of the head as a measurement of where that person is attending. This orientation direction will be used later to generate joint reference behaviors.

- *Chapter 9* : A Simple Mechanism for Social Learning

  Using the basic sensorimotor behaviors and the perceptual system, a mechanism is constructed that allows the robot to mimic the movement of objects or people. For example, if you wave to the robot, the robot will move its arm back and forth in response. Animate trajectories serve as the basic perceptual input which are mapped directly to arm movements.

- *Chapter 10* : Shared Attention Mechanisms

  The second special property of animate agents is that they can have a focus of attention. This focus is a critical social marker, and the robot will be able to respond to that information. Using head posture as an indicator of attention, the robot can engage in shared attention using a feedback mechanism to the existing attention system. These attentional states are also used as a selection mechanism for the mimicry behavior.

- *Chapter 11* : Detecting Simple Forms of Intent

  The final special property of animate agents discussed in this work is that they can maintain an intentional state. Only animate objects can want something or fear something. An additional level of processing is be performed on pairs of trajectories to determine basic representations of intent. Intentional states of attraction and repulsion are classified using a set of motion criteria. These intentional states can be used directly to drive behaviors including reaching or mimicry.

- *Chapter 12* : Toward a Theory of Mind

  The final chapter re-examines the implementation of the embodied theory of mind and discusses extensions to systems that can attribute more complex forms of intention as well as beliefs and desires. This chapter also discusses implications of this model on predictions of human performance and possible extensions that use a robotic implementation as a test-bed for evaluating cognitive models.

Evaluations of individual components are performed throughout the chapters using both comparisons to human performance on similar tasks and subjective evaluations of the performance of the system in responding to naive instructors. Although the systems presented here will not begin to approach some of the complex social skills that children master even in the first year of life, we hope that these foundational skills mark the next step toward the grand challenge of social robots.

Chapter 2

Methodology

> *Because we do not understand the brain very well we are constantly tempted to use the latest technology as a model for trying to understand it. In my childhood we were always assured that the brain was a telephone switchboard. ('What else could it be?') I was amused to see that Sherrington, the great British neuroscientist, thought that the brain worked like a telegraph system. Freud often compared the brain to hydraulic and electro-magnetic systems. Leibniz compared it to a mill, and I am told some of the ancient Greeks thought the brain functions like a catapult. At present, obviously, the metaphor is the digital computer.* – Searle (1986, p. 44)

In the previous chapter, two long-term goals of this research were outlined: to build systems that use natural human social cues to learn from social situations and to evaluate models of human social development using a robotic implementation. These problems are inherently under-specified; our knowledge both of how best to build systems that respond to a variety of social situations and how humans and other animals have evolved to respond to these challenges are not well understood. In fact, even the problem domains are poorly specified. There are many different kinds of social situations and responses, and as an engineering necessity any implemented system will account for only a subset of these possibilities. Even with a restricted class of problem domains, the number of possible solutions is still very large. For example, if the restricted problem domain involves determining whether or not someone is paying attention to you, there are many possible approaches that could perform this task. These approaches could be applied to different behavioral cues (such as head orientation, gaze direction, or posture) and be implemented using different sensory technologies (visible light cameras, infrared cameras, laser range finders, etc.).

This chapter presents some of the methodological principles that have guided the construction of our robotic systems and the implementation of the biological models of social development. We begin with a discussion of the requirements that our two goals introduce. Section 2.1 presents the task requirements for building complex, socially responsive systems and the requirements imposed by attempts to evaluate biological models using a mechanical system. These two sets of requirements leads to a re-evaluation of the driving methodologies in artificial intelligence and robotics research and the questionable assumptions they make about human intelligence (section 2.2). From these observations, we construct a methodology (section 2.3) based upon a modern awareness of cognitive science, neuroscience, psychophysics, and physiology. Many of the ideas represented in this chapter have been published previously in partial form (Brooks & Stein, 1994; Brooks et al., 1998; Scassellati, 1999*a*; Adams et al., 2000).

2.1 Requirements for Building Social Robots

To achieve the goal of building systems that can engage in social learning, a set of minimal requirements must be met. Perhaps the most critical requirement, and the most difficult to satisfy, is that the system must respond in real time and with low latency. A social interaction that is delayed by seconds becomes difficult or even impossible to comprehend. Sensory signals must be processed quickly for relevant social cues, an appropriate behavioral response must be selected, and the motor system must begin executing that behavior, all within a fraction of a second. The latency of the entire system must be very low while the throughput must be very high, which leads to control systems that have relatively few stages of sequential processing.

Social robots will also need to exist in the same complex, noisy, and cluttered environment which people inhabit. The robot must have sophisticated sensing systems to deal with the complexities of the natural environment without artificial simplifications such as static backgrounds, highly engineered workspaces, or restrictions on the shape or coloring of objects in the world. Furthermore, the robot must also interact safely with people and objects in the environment. The robot's control systems must be powerful enough to perform routine tasks (such as lifting small objects), but must incorporate multiple levels of safety protocols.

Our robots must recognize the appropriate social cues for an instructional situation. Complex social scenes require perceptual systems that can extract relevant and salient features in flexible ways. Social signals are inherently multimodal, having visual, auditory, tactile, and other sensory components. To build perceptual systems of this complexity, it will be necessary to use the appropriate mode of sensory processing. The perceptual system must be robust to large variations in the environment, the instructor, and in the dynamic qualities of the instructional content. While current systems will not be capable of understanding all of the rich complexities of human social cues, a successful system must capitalize on the basic social structures that are most significant and that are invariant across individuals.

In addition to recognizing social cues, a social robot must also be able to produce the appropriate social cues that the instructor requires in order to interpret the robot's behavior and to maintain the interaction. Simple gestures such as head nods as well as social indicators such as gaze direction and orientation will allow the instructor to understand the robot's current goals and to correctly interpret the robot's actions. The robot's physical design must also support these behavioral interpretations. This does not imply that the robot must look *exactly* like a human; people have little trouble in interpreting the behavior and expressions of dogs, chimpanzees, and other animals. The robot must, however, have a sufficient appearance that the human instructor can easily apply his knowledge of social responses to the robot's behavior.

In recognizing and exhibiting social cues, it is critical that the behavior and appearance of the robot provide an appropriate level of expectation in the instructor. People become quickly frustrated when they are faced with machines or interfaces that appear to provide more functionality than the machine actually can exhibit (Norman, 1990). At the same time, it is also important that people be able to apply the same types of social assumptions to our robots that they would to a person in a similar circumstance. The robot must not be so limiting in its interface or appearance that the human instructor cannot use normal social proficiencies to understand the robot. This will naturally result in people attributing qualities such as intention, feelings, and desires to the robot. The robot's design must facilitate such attributions without providing expectations that are too extravagant.

There are certainly many other design criteria that a social machine must meet in addition to the typical engineering criteria of cost, reliability, robustness, and availability. Design criteria for social constructs have been treated more thoroughly by other authors (Dautenhahn, 1997; Breazeal, 2000) and any good engineering design book can give general pointers for evaluating standard design criteria. However, one further design criterion deserves mention here. Socially adept systems must respond based on the "hidden" states of goal, desire, and intent rather than on explicit actions. Just as human infants respond to the intention of an action rather than the action itself (Meltzoff, 1995), our robotic systems should ideally respond to the intent of the instructor rather than the explicit actions of the instructor. In other words, these socially intelligent machines need a theory of mind.

2.1.1 Classical Approaches in Artificial Intelligence

Many researchers in artificial intelligence have also attempted to satisfy subsets of these design criteria by building systems that in some way model the only universally accepted example of intelligence: humans. However, many of these classical approaches have been based upon either introspective analysis of the respective authors or on an understanding of human psychology and neurology that is outdated (Brooks et al., 1998). This section considers some of these classical approaches as a background against which to view the alternative methodologies used in this work. While we will also certainly commit similar errors in constructing a new methodology, it is an untenable position to remain locked in the past.

One of the most basic methodological models in artificial intelligence was the claim of Newell & Simon (1961) that humans use physical symbol systems to "think". Over time, this has become adopted into artificial intelligence as an implicit and dominant hypothesis (see Brooks (1991*a*) for a review). Following this approach, an AI system would generally rely on uniform, explicit, internal representations of capabilities of the system, the state of the outside world, and the desired goals. These AI systems are concerned primarily with search problems to both access the relevant facts and to determine how to apply them. More recent approaches incorporate Bayesian or other probabilistic ideas into this basic framework (Pearl, 1988). These neo-classical approaches explicitly represent and manipulate probabilities.

We believe that classical and neo-classical AI falter in assuming that because a description of reasoning/behavior/learning is possible at some level, then that description must be made explicit and internal to any system that carries out the reasoning/behavior/learning. The realization that descriptions and mechanisms could be separated was one of the great breakthroughs of Rosenschein & Kaelbling (1986), but unfortunately that realization has been largely ignored. This introspective confusion between surface observations and deep structure has led AI away from its original goals of building complex, versatile, intelligent systems and towards the construction of systems capable of performing only within limited problem domains and in extremely constrained environmental conditions. While these classical techniques may be useful tools in other domains, they are inappropriate for building perceptually-grounded systems.

The next section of this chapter explores the assumptions about human intelligence which are deeply embedded within classical AI. The following sections explain a methodology which rectifies these mistakes and yields an alternate approach to creating robustly functioning intelligent systems.

2.2 Questionable Assumptions about Intelligence

Technological advances have always influenced the metaphors and models that we have used to understand ourselves. From ancient Greece and Rome through the dark ages, the first marvels of chemistry (perhaps better thought of as alchemy in that era) were beginning to take shape. Philosophers at that time spoke of the different humours in each person that must remain balanced, just as the formulations of an elixir required exact proportions of different ingredients (Gleitman, 1991). In the early 1900's, the dominant technological wonder was the steam engine in its many forms. Freud and other psychologists of this era often spoke about the brain as if it were a hydraulic system; the pressures building within the ego and id could be released (as if a valve had been turned), but left unattended would build in intensity until an explosion occurred (Freud, 1962). Throughout the twentieth century, the metaphors changed rapidly from hydraulic systems to basic electronic systems such as the telegraph and the telephone switchboard until they naturally landed upon the digital computer. Today, discussions of memory capacity and storage methods, processing power, and information flow are as likely to occur in a psychology lecture as in a computer science lecture.

Early artificial intelligence systems exploited the computational metaphors of mind in an attempt to explain human behavior. While these classical approaches were certainly a necessary step, in a way it became too easy to follow the metaphor by using the technology on which the metaphor was based. The biases that the computational metaphor of mind introduced have carried over into large portions of artificial intelligence research. These biases, combined with an emphasis on subjective observation and introspection, have lead to a number of conceptual errors in how artificial intelligence (and some other parts of the cognitive sciences) have come to think about human intelligence. The metaphor has been taken too far (Brooks, 1991*a*,*b*).

Three of these conceptual errors have been particularly damaging: presuming the presence of monolithic internal models, monolithic control, and general purpose processing. These assumptions result from a reliance on the computational metaphors (such as mathematical logic, Von Neumann architectures, etc.), but are refuted by a modern understanding of cognitive science and neuroscience.

2.2.1 No Full Monolithic Internal Models

In classical artificial intelligence, sensory perceptions are used to construct a single, consistent internal model of the external world. This internal model is often a three-dimensional representation of the external environment, or alternately a purely abstract representation filled with logical relations. All incoming sensory data is converted into a format that can be processed in this internal model. The job of the perceptual system is to convert complex sensory signals into representational states within this model. This single monolithic model is attractive from an introspective stance because we see ourselves as existing in a primarily static world; I can shut my eyes and see the room that I have been sitting in, and I can think and plan my next actions by manipulating that model. This monolithic internal representational scheme that somehow mirrors the external world has been attacked from multiple directions in psychology, psychophysics, and neurology.

There is evidence that in normal tasks humans tend to minimize their internal representation of the world. Ballard et al. (1995) have shown that in performing a complex task, like building a copy of a display of blocks, humans do not normally maintain an internal model of the entire visible scene. Subjects in their experiments were asked to copy a demonstration structure (the model) in an empty workspace using the same components. Subjects would initially look at the model, and then shift their gaze and their attention to the workspace, return to the model, and repeat. By changing the model display while subjects were looking at the workspace, Ballard found that subjects noticed only the most drastic of changes; rather than keeping a complete model of the scene, they instead left that information in the world and continued to refer back to the scene while performing the copying task. If subjects had been maintaining internal models of the demonstration structure, then they should have been able to notice these drastic changes. Similar results have been seen in the psychophysics community in the areas of change blindness (Rensink et al., 1997) (in which large portions of the visual scene are changed either during eye movements or when accompanied by a flashing display) and inattentional blindness (Mack & Rock, 1998) (in which observers fail to notice objects or events to which they are not attending).

There is also evidence that there are multiple internal sensory or logical representations, which are not mutually consistent. For example, in the phenomena of blindsight, cortically blind patients can discriminate different visual stimuli, but actually report seeing nothing (Weiskrantz, 1986). These patients report having no visual experiences within some portion of their visual field, and yet at times can perform tasks, such as orienting a piece of mail properly to enter either a vertical or horizontal slot, that rely on that part of the visual field. Some of these subjects are also significantly better than chance when forced to make choices involving the affected visual area. These inconsistencies would not be a feature of a single central model of visual space.

These experiments and others like it (e.g., Gazzaniga & LeDoux, 1978) demonstrate that humans do not construct a full, monolithic model of the environment. Instead, we tend to only represent what is immediately relevant from the environment, and those representations do not have full access to one another.

2.2.2 No Monolithic Control

Naive introspection and observation can lead one to believe in a neurological equivalent of the central processing unit – something that makes the decisions and controls the other functions of the organism. While there are undoubtedly control structures, this model of a single, unitary control system is not supported by evidence from cognitive science.

One example comes from studies of split brain patients by Gazzaniga & LeDoux (1978). These are patients where the corpus callosum (the main structure connecting the two hemispheres of the brain) has been cut.[1] The patients are surprisingly normal after the operation, able to resume their normal lives after a recovery period. By careful experimental procedures, Gazzaniga and LeDoux were able to isolate the deficits caused by this procedure by presenting different information to either side of the (now unconnected) brain. Since each hemisphere controls one side of the body, the experimenters could probe the behavior of each hemisphere independently (for example, by observing the subject picking up an object appropriate to the scene that they had seen). In perhaps the most well-known example, a snow scene was presented to the right hemisphere and the leg of a chicken to the left hemisphere of a teenage male subject. The subject was asked to select two items (one with each hand) from a set of binocularly viewed objects based on the scene that he had just seen. The subject selected a chicken head to match the chicken leg, explaining with the verbally dominant left hemisphere that "I saw the claw and picked the chicken". When the right hemisphere then picked a shovel to correctly match the snow, the left hemisphere explained that you need a shovel to "clean out the chicken shed" (Gazzaniga & LeDoux, 1978, p.148). The separate halves of the subject's brain independently acted appropriately, but one side falsely explained the choice of the other. This suggests that there are multiple independent control systems, rather than

[1] This somewhat radical procedure was originally attempted as a last-resort treatment for severe epilepsy. The hope was that because the two hemispheres of the brain would be electrically isolated, a seizure would effect a smaller portion of the body and be less disruptive and dangerous. The treatment was remarkably successful.

a single monolithic one. This experiment can also be conducted on normal subjects by injecting sodium amytal into the carotid artery, which effectively anesthetizes one hemisphere. Stimuli can then easily be presented to only one hemisphere and testing can occur either immediately or after the anesthesia wears off. These experiments force us to recognize that humans are capable of holding conflicting and non-consistent beliefs.

2.2.3 Not General Purpose

The brain is conventionally thought to be a general purpose machine, acting with equal skill on any type of operation that it performs by invoking a set of powerful rules. However, humans seem to be proficient only in particular sets of skills, at the expense of other skills, often in non-obvious ways. A good example of this is the Stroop effect (Stroop, 1935). In this simple task, subjects are asked to read a column of words as quickly as they can. Each of the words is the name of a color (e.g., "blue," "red," or "yellow") which is printed in an ink color that does not match the word's meaning. For example, the first instance of the word "blue" might be printed in red ink, while the word "red" might be printed in yellow ink. Subjects produce more errors, and are much slower, at reading this list when the ink colors do not match the labels than when the colors do match the labels. Performance in this color recognition and articulation task is actually dependent on the semantic content of the words. If our capacity for reading were truly a general purpose process, why would altering the color of the ink effect performance? This experiment demonstrates the specialized nature of human computational processes and interactions. Similar perceptual cross-over effects can be seen between vision and audition (Churchland et al., 1994) and vice-versa (Cohen & Massaro, 1990).

One might complain that the Stroop effect is purely a perceptual event, and that while perceptual processing may contain domain specific mechanisms, the more cerebral functions of deductive logic, planning, and behavior selection are general purpose. However, the evidence from psychology also refutes this hypothesis. For example, in the area of deductive logic, humans often perform extremely poorly in different contexts. Wason (1966) found that subjects were unable to apply the negative rule of if-then inference when four cards were labeled with single letters and digits. Subjects were shown four cards, each of which contained a letter written on one side and a number written on the reverse. The cards were placed flat on a table, so the observer saw only one side of the card. For example, the subject might see cards that said "E", "F", "4", and "7". The task for this presentation to determine which cards *must* be flipped over to verify whether or not the following rule held true: if a card has a vowel on one side, then there is an even number on the other side. Only 4% of the university student subjects in the original study correctly answered "E" and "7".[2] However, when subjects were given a task with the same logical structure but

[2] For the rule to be true, the "E" must have an even number on the other side and the "7" must not have a vowel on the other side.

a different context, they had little difficulty in applying the correct logical inferences. For example, if the cards contained a person's age on one side and their choice of beverage on the other, the set "Gin", "Soda", "22", "16" has the same logical structure when attempting to verify the rule: if a person is drinking alcohol, they must be at least 21 years of age. Similar populations of university students had no difficulty in solving this task (Griggs & Cox, 1982).

Further, humans often do not use subroutine-like rules for making decisions. They are often more emotional than rational, and there is evidence that this emotional content is an important aspect of decision making (Damasio, 1994). For example, Damasio's patient Elliot suffered from a meningioma located above the nasal cavities which compressed and destroyed large portions of the frontal lobe bilaterally. After the surgical removal of the tumor, Elliot had no permanent paralysis but had a notable lack of affective response. Elliot reported having very few emotional feelings. While Elliot had normal sensory abilities and normal motor responses, his decision making skills were severely impaired. Elliot had problems in making normal judgments that people make continuously throughout their normal daily activities (such as what to order for lunch) and in evaluating whether a particular decision was risky or safe. Damasio has proposed that one purpose of emotional responses is to quickly label possible actions as good or bad and to expedite the decision making process by guiding the selection of one possible course of action among many. Damasio has named this the somatic marker hypothesis to indicate that these emotional responses place a "marker" on each of the possible decision points that guides selection. Damasio argues that the evidence from Elliot and patients like him fail to support models of decision making based solely on abstract logic or propositional statements.

2.3 Alternate Essences of Human Intelligence

Humans have the ability to autonomously learn, generalize, organize, and assimilate immense numbers of skills, competencies and facts. We believe that these abilities are a direct result of four intertwined key human attributes:

- *Development* forms the framework by which humans successfully acquire increasingly more complex skills and competencies.

- *Social interaction* allows humans to exploit other humans for assistance, teaching, and knowledge.

- *Embodiment and physical coupling* allow humans to use the world itself as a tool for organizing and manipulating knowledge.

- *Integration* allows humans to maximize the efficacy and accuracy of complementary sensory and motor systems.

Since humans are vastly complex systems, we do not expect to duplicate every facet of human intelligence. However, we must be very careful not to ignore aspects of human intelligence solely because they appear complex. Classical and neo-classical AI tend to ignore or avoid these complexities, in an attempt to simplify the problem (Minsky & Papert, 1970). We believe that many of these discarded elements are essential to human intelligence and that they actually simplify the problem of creating human-like intelligence.

2.3.1 Development

Humans are not born with complete reasoning systems, complete motor systems, or even complete sensory systems. Instead, they undergo a process of development where they are able to perform more difficult tasks in more complex environments en route to the adult state. This is a gradual process, in which earlier forms of behavior disappear or are modified into more complex types of behavior. The adaptive advantage of the earlier forms appears to be that they prepare and enable more advanced forms of behavior to develop within the situated context they provide. The developmental psychology literature abounds with examples of this phenomenon. For instance, the work of Diamond (1990) shows that infants between five and twelve months of age progress through a number of distinct phases in the development of visually guided reaching. In one reaching task, the infant must retrieve a toy from inside a transparent box with only one open side. In this progression, infants in later phases consistently demonstrate more sophisticated reaching strategies to retrieve the toy in more challenging scenarios. As the infant's reaching competency develops, later stages incrementally improve upon the competency afforded by the previous stage.

Building systems developmentally facilitates learning both by providing a structured decomposition of skills and by gradually increasing the complexity of the task to match the competency of the system. The developmental process, starting with a simple system that gradually becomes more complex, allows efficient learning throughout the whole process. For example, infants are born with low-acuity vision. The infant's visual performance develops in step with their ability to process the influx of stimulation (Johnson, 1993). The same is true for the motor system. Newborn infants do not have independent control over each degree of freedom of their limbs, but through a gradual increase in the granularity of their motor control they learn to coordinate the full complexity of their bodies. A process where the acuity of both sensory and motor systems are gradually increased significantly reduces the difficulty of the learning problem (Thelen & Smith, 1994).

To further facilitate learning, the gradual increase in internal complexity associated with development should be accompanied by a gradual increase in the complexity of the external world. For an infant, the caregiver biases how learning proceeds by carefully structuring and controlling the complexity of the environment. This approach is in stark contrast to most machine learning methods, where the robot learns in a usually hostile environment, and the bias, instead of coming from the robot's interaction with the world, is included by the designer. We believe that gradually increasing the complexity of the environment makes learning easier and more robust.

By exploiting a gradual increase in complexity both internal and external, while reusing structures and information gained from previously learned behaviors, we hope to be able to learn increasingly sophisticated behaviors. We believe that these methods will allow us to construct systems which do scale autonomously (Ferrell & Kemp, 1996).

2.3.2 Social Interaction

Human infants are extremely dependent on their caregivers, relying upon them not only for basic necessities but also as a guide to their development. The presence of a caregiver to nurture the child as it grows is essential. This reliance on social contact is so integrated into our species that it is hard to imagine a completely asocial human. However, severe developmental disorders sometimes give us a glimpse of the importance of social contact. One example is autism (DSM, 1994; ICD, 1993; Cohen & Volkmar, 1997). Autistic children often appear completely normal on first examination; they look normal, have good motor control, and seem to have normal perceptual abilities. However, their behavior is completely strange to us, in part because they do not recognize or respond to normal social cues (Baron-Cohen, 1995). They do not maintain eye contact, recognize pointing gestures, or understand simple social conventions. Even the most highly functioning autistics are severely disabled in our society. Building social skills into an artificial intelligence provides not only a natural means of human-machine interaction but also a mechanism for bootstrapping more complex behavior.

Social interaction can be a means to facilitate learning. New skills may be socially transfered from caregiver to infant through mimicry or imitation, through direct tutelage, or by means of scaffolding, in which a more able adult manipulates the infant's interactions with the environment to foster novel abilities. Commonly scaffolding involves reducing distractions, marking the task's critical attributes, reducing the number of degrees of freedom in the target task, and enabling the subject to experience the end or outcome before the infant is cognitively or physically able of seeking and attaining it for herself (Wood et al., 1976).

2.3.3 Embodiment and Physical Coupling

Perhaps the most obvious, and most overlooked, aspect of human intelligence is that it is embodied. Humans are embedded in a complex, noisy, constantly changing environment. There is a direct physical coupling between action and perception, without the need for an intermediary representation. This coupling makes some tasks simple and other tasks more complex. By exploiting the properties of the complete system, certain seemingly complex tasks can be made computationally simple. For example, when putting a jug of milk in the refrigerator, you can exploit the pendulum action of your arm to move the milk (Greene, 1982). The swing of the jug does not need to be explicitly planned or controlled, since it is the natural behavior of the system. Instead of having to plan the whole motion, the system only has to modulate, guide and correct the natural dynamics. For an embodied system, internal representations can be ultimately grounded in sensory-motor interactions with the world (Lakoff, 1987).

A principle tenet of our methodology is to build and test real robotic systems. We believe that building human-like intelligence requires human-like interaction with the world (Brooks & Stein, 1994). Humanoid form is important to allow humans to interact with the robot in a natural way. In addition, we believe that building a real system is computationally less complex than simulating such a system. The effects of gravity, friction, and natural human interaction are obtained for free, without any computation.

Another aspect of our methodology is to exploit interaction and tight coupling between the robot and its environment to give complex behavior, to facilitate learning, and to avoid the use of explicit models. Our systems are physically coupled with the world and operate directly in that world without any explicit representations of it (Brooks, 1986, 1991*b*). There are representations, or accumulations of state, but these only refer to the internal workings of the system; they are meaningless without interaction with the outside world. The embedding of the system within the world enables the internal accumulations of state to provide useful behavior (this was the fundamental approach taken by Ashby (1960) contemporaneously with the development of early AI).

2.3.4 Integration

Humans have the capability to receive an enormous amount of information from the world. Visual, auditory, somatosensory, and olfactory cues are all processed simultaneously to provide us with our view of the world. However, there is evidence that the sensory modalities are not independent; stimuli from one modality can and do influence the perception of stimuli in another modality. Churchland et al. (1994) describe an experiment illustrating how audition can cause illusory visual motion. A fixed square and a dot located to its left are presented to the observer. Without any sound stimuli, the blinking of the dot does not result in any perception of motion. If a tone is alternately played in the left and right ears, with the left ear tone coinciding with the dot presentation, there is an illusory perception of back and forth motion of the dot, with the square acting as a visual occluder. Vision can cause auditory illusions too, for example in the McGurk effect (Cohen & Massaro, 1990). These studies demonstrate that humans' perception of their senses cannot be treated as completely independent processes.

Some tasks are best suited for particular sensory modalities. Attempting to perform the task using a different modality is sometimes awkward and computationally intensive. Using the appropriate mode for a given task can reduce the requisite computation. By integrating multiple sensory and motor systems, we can address a wider range of tasks while maintaining computational feasibility.

By integrating different sensory modalities we can exploit the complex nature of stimuli to facilitate learning. For example, objects that make noise often move. This correlation can be exploited to facilitate perception. These relationships have been extensively characterized for the case of the development of auditory localization. Wertheimer (1961) has shown that vision and audition interact from birth; even ten-minute-old children will turn their eyes toward an auditory cue. Related investigations with young owls have determined that visual stimuli greatly affect the development of sound localization. With a constant visual bias from prisms worn over the eyes, owls adjusted their sound localization to match the induced visual errors (Knudsen & Knudsen, 1985).

2.4 Conclusion

Building systems that can both solve interesting and challenging social tasks and also serve as a testbed for evaluating models of social development in children introduces a set of difficult constraints. The system must operate in real time, recognizing the appropriate social cues from the human instructor and providing social cues in response that are easy and natural to interpret. The robot must be appealing to interact with, must be easy to attribute intentions and goals to, and yet must not appear to be capable of more than it can accomplish.

In addressing these issues, we have moved away from the areas of classical AI and the accompanying computational metaphors of mind. We have thus chosen to approach AI from a different perspective, in the questions we ask, the problems we try to solve, and the methodology and inspiration we use to achieve our goals. By examining a more thorough understanding of current research in human psychology, neurology, and psychophysics, we have examined and rejected many of the common assumptions that the computational metaphor produces. Further, we have proposed a set of four characteristics which serve as the core to our methodology in constructing these systems. The principles of development, social interaction, physical coupling to the environment, and integration will be essential to guide us towards our goal.

Chapter 3

Developmental Models of Theory of Mind

> *An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory because such states are not directly observable, and the system can be used to make predictions about the behaviour of others.* – Premack & Woodruff (1978, p. 516)

Research from many different disciplines has focused on theory of mind. Students of philosophy have been interested in the understanding of other minds and the representation of knowledge in others. Most recently, Dennett (1987) has focused on how organisms naturally adopt an "intentional stance" and interpret the behaviors of others as if they possess goals, intents, and beliefs. Ethologists have also focused on the issues of theory of mind. Studies of the social skills present in primates and other animals have revolved around the extent to which other species are able to interpret the behavior of conspecifics and influence that behavior through deception (e.g., Premack, 1988; Povinelli & Preuss, 1995; Cheney & Seyfarth, 1991). Research on the development of social skills in children has focused on characterizing the developmental progression of social abilities (e.g., Fodor, 1992; Wimmer & Perner, 1983; Frith & Frith, 1999) and on how these skills result in conceptual changes and the representational capacities of infants (e.g., Carey, 1999; Gelman, 1990). Furthermore, research on pervasive developmental disorders such as autism has focused on the selective impairment of these social skills (e.g., Perner & Lang, 1999; Karmiloff-Smith et al., 1995; Mundy & Sigman, 1989).

This chapter will review some of the basic observations concerning the set of skills collectively called "theory of mind" (section 3.1). We then present two of the most popular and influential models, one from Leslie (1994) (section 3.2) and one from Baron-Cohen (1995) (section 3.3), which attempt to link together multidisciplinary research into a coherent developmental explanation. Section 3.4 will discuss the implications of these models for the construction of humanoid robots that engage in natural human social dynamics and will also highlight some of the issues involved in implementing the structures that these models propose. Finally, Section 3.5 will describe a hybrid model called *embodied theory of mind* that links together ideas from both Baron-Cohen and Leslie with a grounded perceptual system. I do not mean to imply by this choice of names that the other models exist in a vacuum without reference to actual physical behavior. However, the differences in the hybrid model came about as a direct result of attempts to implement these basic foundational skills on an embodied robotic system. The hybrid model serves as the basis for an implementation for a humanoid robot that will be discussed in the following chapters.

3.1 Basic observations

The term "theory of mind" has been used to identify a collection of socially-mediated skills which are useful in relating the individual's behavior within a social context.[1] Examples of these skills include detecting eye contact, recognizing what someone else is looking at, pointing to direct attention to interesting objects, and understanding that other people have ideas that differ from one's own. The most important finding regarding these skills, repeated in many different forms, is that "theory of mind" is not a single monolithic system. Evidence from childhood development shows that not all of these skills are present from birth, and there is a stereotypic progression of skills that occurs in all infants at roughly the same rate (Hobson, 1993). Children master certain skills (such as recognizing when someone is making eye contact with them) before acquiring more complex skills (such as pointing to desired objects).

A second perspective on this decomposition can be seen in the presence of these same skills in non-human animals. The same ontogenetic progression of 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 closer and closer to humans on a phylogenetic scale (Povinelli & Preuss, 1995). Abilities that most six-month-old human children have mastered are found in many vertebrates, while skills characteristic of a child at 15 months are found only in primates.

Finally, 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

[1] Other authors have attempted to distinguish between the "theory of mind skills" and certain pre-cursor abilities such as maintenance of eye contact (Mundy & Sigman, 1989). For the work presented here, this difference is largely irrelevant, and the term "theory of mind skills" will include these precursor abilities.

the critical deficit may be a lack of advanced theory of mind skills (Baron-Cohen, 1995). In comparison to other mental retardation and developmental disorders (like Williams and Down's Syndromes), the social deficiencies of autism are quite specific (Karmiloff-Smith et al., 1995).

The simplest theory of mind skills are those that emerge earliest in a child's development, are most likely to be intact in autistic individuals, and are found in a wide variety of animals. The more complex skills are the last to be observed developmentally, are the most likely to be impaired in autism, and are observed only in humans. In this section, we describe details of many of the social skills that are discussed in studies of theory of mind. The following section describes two models that attempt to integrate these behavioral findings into comprehensive explanations of the development and acquisition of these skills.

3.1.1 Eye Contact

One of the most basic social skills 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*b*), 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 (Fagan, 1976).

A slightly more complex behavior is gaze following, which 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).[2] 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 (Butterworth, 1991). Gaze following is a joint attention mechanism, that is, it 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 et al., 1976).

[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.

3.1.2 Pointing

A second set of behaviors involves pointing. Developmental psychologists often distinguish between *imperative* pointing and *declarative* pointing. Imperative pointing is a gesture used to obtain an object that is out of reach by extending the arm toward that object. This behavior is first seen in human children at about nine months of age (Baron-Cohen, 1995) and has been observed in many primates (Cheney & Seyfarth, 1990). However, there is nothing particular to this behavior that is different from a simple reach; the infant is initially as likely to perform imperative pointing when the adult is attending to the infant as when the adult is looking in the other direction or when the adult is not present. The adult's interpretation of the 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 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 adult).

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 outside their reach, such as the sun or an airplane passing overhead, or to objects that they have no interest in approaching, such as an angry dog. Declarative pointing also only occurs under specific social conditions. Children do not point unless there is someone to observe their action and often use other social conventions to draw attention to the object of interest. No other species has been shown to be responsive to declarative points and to generate declarative points in naturalistic circumstances (Povinelli & Preuss, 1995).

3.1.3 Responding to Intent

Theory of mind abilities often bridge the gap between high-level cognitive properties and low-level perceptual properties. For example, the attribution of intention to an object is often characterized as a complex, high-level cognitive task involving reasoning and episodic memory. Many theory of mind models ground these cognitive properties in specific low-level percepts. Heider & Simmel (1944) were the first to characterize the basic perceptual nature of intentional attribution (figure 3-1). Subjects in their experiments were shown movies of simple geometric objects moving against a static background and were asked to describe the content of the movies. In spontaneous utterances, virtually all subjects used words like "wants," "fears," or "needs," in describing the movement of these geometric shapes. The anthropomorphization of these shapes was completely automatic; subjects found it extremely difficult to describe the scene purely in geometric terms even when prompted to do so. Heider and Simmel's original results have been replicated and enhanced in many ways (for a review, see Scholl & Tremoulet, 2000), but the basic observation that humans

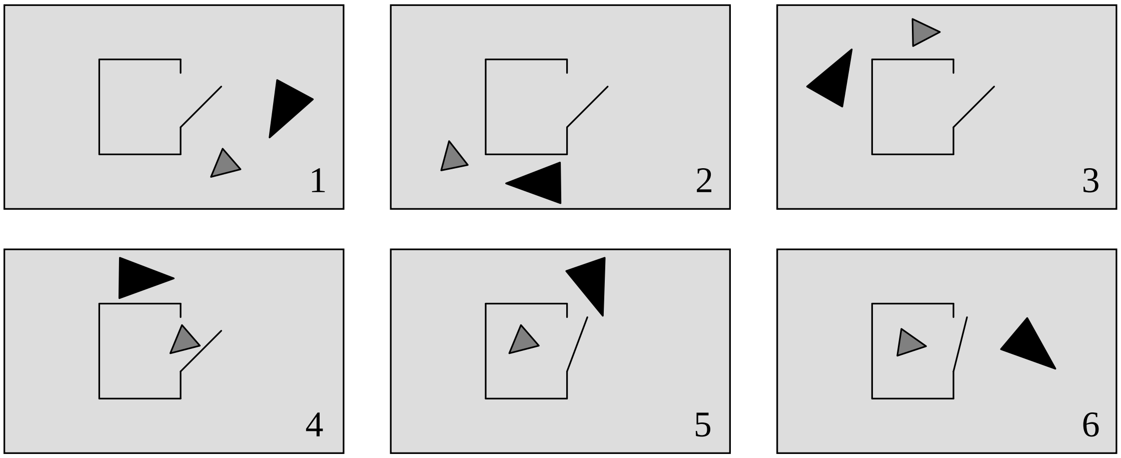
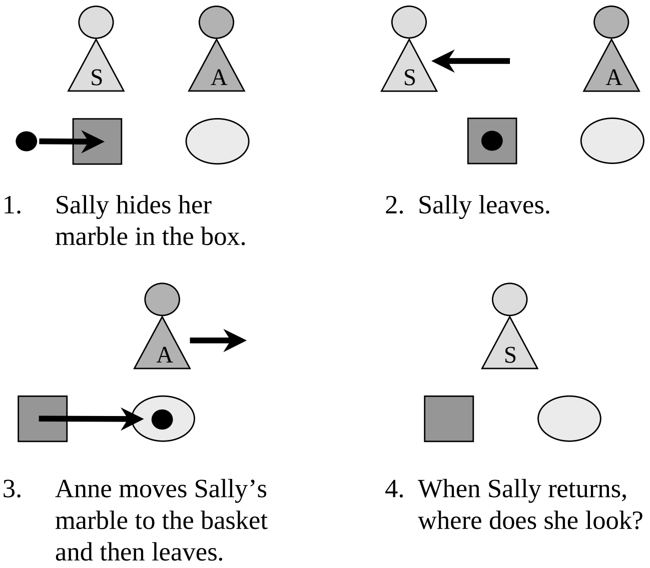


Figure 3-1: Six frames from a movie sequence similar to those used by Heider & Simmel (1944). Subjects readily attribute intent and goal to the movements of even these simple geometric shapes. In this example, the large triangle is seen as *wanting* to catch the smaller triangle.

naturally tend to attribute intentional states to even simple minimalistic perceptual scenes remains unchallenged.

Meltzoff (1995) has demonstrated that infants as young as 18 months of age are also sensitive to the intent of an action and are capable of acting based on the desired outcome of an unsuccessful intentional act. Infants of this age who were shown an adult pulling apart a toy shaped like a dumbbell had no difficulty in reproducing the action. When infants were shown the adult attempting to perform the same action but failing when their fingers slipped off the object, the infants tended to respond by completing the action and pulling the object apart rather than imitating the slipping motion exactly. Interestingly, infants in this study failed to imitate the intended act when a mechanical set of pincers replaced the human hands. The attribution of intention in this case was contingent on the nature of the agent that performed the action.

This sensitivity to intent is also seen in many parts of the animal kingdom. The simplest forms of intent are expressions of desire and fear, which can easily be expressed in terms of predator and prey relationships. The evolutionary advantage of this discrimination power is easy to see; the animal that can determine whether it is being chased (or can detect the animal that fears it) has an obvious selective advantage. In some primates, the evolutionary arms race has resulted in behaviors that attempt to deceive conspecifics by masking the animal's true intentions (Byrne & Whiten, 1988; Whiten & Byrne, 1997). Many animal cultures are based on dominance hierarchies that are established not through explicit combat but rather through complex threat displays and responses. The interpretation of these actions by other observers can be demonstrated since other individuals in the social group learn dominance relations by observing these interactions (Cheney & Seyfarth, 1990).



1. Sally hides her marble in the box.

2. Sally leaves.

3. Anne moves Sally's marble to the basket and then leaves.

4. When Sally returns, where does she look?

Figure 3-2: The Sally-Anne test of false belief. See text for description. Adapted from Baron-Cohen et al. (1985).

3.1.4 False Belief Tasks

Perhaps the most widely recognized test of theory of mind has been the false belief task. This class of experimental designs focuses on the question of whether or not a child can represent that an external agent maintains a belief that is different from the child's own beliefs. In many ways, these tasks are more difficult than any of the preceding problems, as they require the use of many of these precursor abilities and are often reliant on linguistic responses. Performance on these tasks may not be effectively measured on non-verbal children or animals.

The first false belief task to be widely studied was the Sally-Anne task (see figure 3-2) (Baron-Cohen et al., 1985), which was based on an earlier design by Wimmer & Perner (1983). In this scenario, the child is shown a short play that involves two agents, Sally and Anne, and two locations, a basket and a box. Sally enters the room carrying a marble, which she hides inside the box. Sally leaves the room, and while she is away, Anne moves the marble from the box to the basket. Anne covers up the marble inside the basket so that it is not visible and then leaves the room. Sally re-enters the scene and the child is asked where Sally will look first in order to find her marble. To correctly solve this task, children must represent that the belief that they hold (that the marble is in the basket) is different from the belief that Sally holds (that the marble is still in the box). Normal children are able to pass this test at 3-4 years of age, but fail at younger ages (almost always by responding that Sally will look in the location where the marble is actually hidden). Baron-Cohen et al. (1985) tested both individuals with autism and individuals with Down's Syndrome who had a sufficient verbal aptitude of greater than four years of age. Virtually all autistic individuals failed this test while those with Down's Syndrome passed.

The Sally-Anne task has received criticism from many sources (Whiten, 1991), much of which is deserved in that the task requires many competencies and a very complex understanding of the social situation. A simpler false belief task comes from Perner et al. (1989), which has come to be known as the "Smarties" test.[3] In this test, the child is shown a sealed tube of Smarties and asked "What do you think is in here?" Once the child responds that there are Smarties in the container, the experimenter opens the tube to reveal that instead there are pencils inside. The children show surprise (and often disappointment) at the appearance of the pencils. The experimenter then seals the container and asks two belief questions: "Before I opened the tube, what did you think was inside?" and "When the next child comes in (who has not seen inside the tube), what will he think is inside?" Normal children above 3-4 years of age correctly answer "Smarties" to both belief questions. Younger children, and the majority of autistic children, respond to both belief questions with "pencils." In failing this test, the child demonstrates an inability to reconcile the current (true) belief that there are pencils inside the tube with both the prior (false) belief that the child himself held and the naive (false) belief that another child would have in a similar circumstance.

3.2 Leslie's Model

With a wealth of data on the developmental organization of these disparate abilities, there have been two major attempts to organize and explain this data. The first comes from Leslie (1994). Leslie's theory treats the representation of causal events as a central organizing principle to theories of object mechanics and theories of other minds much in the same way that the notion of number may be central to object representation. According to Leslie, the world is naturally decomposed into three classes of events based upon their causal structure: one class for *mechanical agency*, one for *actional agency*, and one for *attitudinal agency*. Leslie argues that evolution has produced independent domain-specific modules to deal with each of these classes of event. The Theory of Body module (ToBY) deals with events that are best described by mechanical agency, that is, they can be explained by the rules of *mechanics*. The second module is system 1 of the Theory of Mind module (ToMM-1) which explains events in terms of the intent and goals of agents, that is, their *actions*. The third module is system 2 of the Theory of Mind module (ToMM-2) which explains events in terms of the *attitudes* and beliefs of agents.

3.2.1 ToBY: The Theory of Body

The Theory of Body mechanism (ToBY) embodies the infant's understanding of physical objects. ToBY's goal is to describe the world in terms of the mechanics of physical objects and the events they enter into. In this sense, ToBY encapsulates a certain

[3] Smarties are a candy common in Great Britain, where these experiments were originally performed. It is a safe assumption that all British children in the study would be able to instantly recognize the characteristic tubular Smarties package.

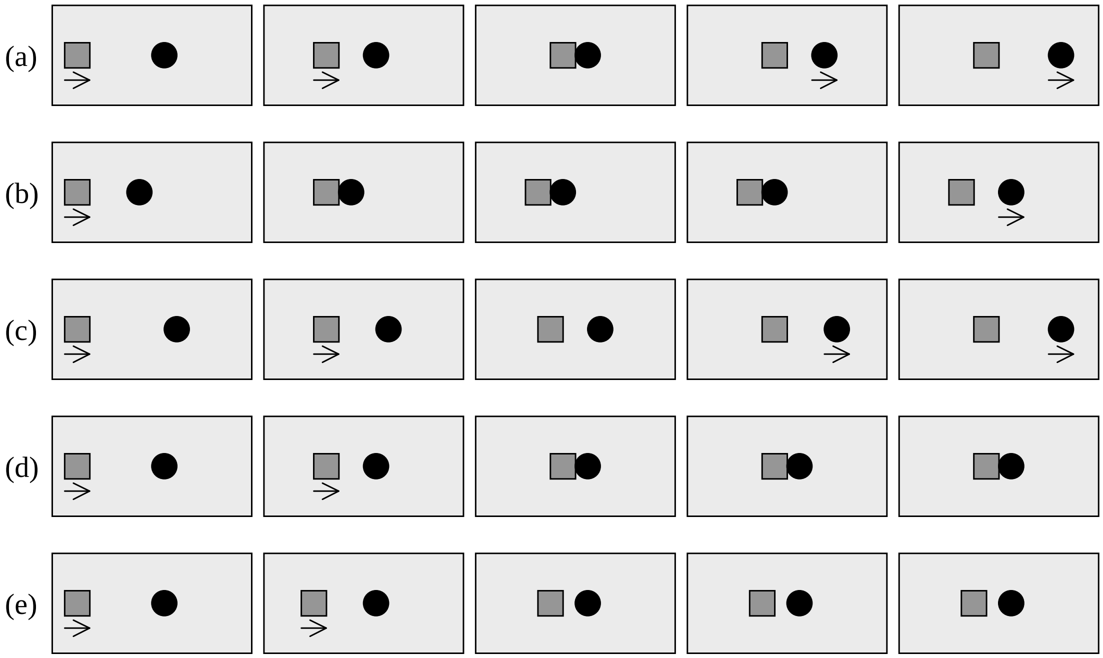


Figure 3-3: Sequences from some of Michotte's basic experiments on perceptual causality. Each row represents five frames from an image sequence involving a gray square and a black circle. In the first sequence (a), the observer has the perception of mechanical causation – the black circle moves because it was struck by the gray square. However, if a temporal gap (b) or a spatial gap (c) is introduced, the movement of the circle is seen as originating from the circle itself. Similarly, cases where contact is made (d) but there is no movement are seen as unusual while cases where no contact is made and no motion results (e) are normal physical processes.

naive understanding of physics. Note that the knowledge in ToBY is neither an accurate view of physics nor is it completely accessible to conscious introspection; ToBY cannot explain how a gyroscope works both because the explanation is not within its explanatory power and because the details of that explanation would be inaccessible to other processes.

ToBY in humans is believed to operate on two types of visual input: a three-dimensional object-centered representation from high level cognitive and visual systems and a simpler motion-based system. This motion-based system accounts for the causal explanations that adults give (and the causal expectations of children) to the "billiard ball" type launching displays pioneered by Michotte (1962) (see figure 3-3). Michotte observed that even with a relatively simple visual stimulus, adult observers were very quick to attribute causal explanations to the movement of simple stimuli. For example, in case (a) of figure 3-3, observers report that the black circle moves because it was struck by the gray square. However, slight alterations of the temporal (b) or spatial (c) characteristics of the collision result in a very different kind of causal explanation. In these cases, observers report that the black circle moves of its own accord, that it "wants" to move. Leslie proposed that this sensitivity to the spatio-temporal properties of events is innate, but more recent work from Cohen & Amsel (1998) may show that it develops extremely rapidly in the first few months and is fully developed by 6-7 months. Cohen and Amsel further argue that infants younger than 6 months respond to spatio-temporal changes in the stimulus but without reference to the causal properties. We will return to the details of this mechanism in chapter 7.

3.2.2 ToMM: The Theory of Mind Mechanism

Just as the theory of body mechanism deals with the physical laws that govern objects, the theory of mind mechanism deals with the psychological laws that govern agents. The objective of ToMM is to interpret the internal state of other agents by making assumptions based on their behavior. These internal states of belief, goal, and desire cannot be observed directly, but rather must be estimated from the actions that the agent takes. The theory of mind mechanism performs this state estimation for two slightly different classes of causal events. The first deals with actional agencies, that is, it explains the actions of social agents in terms of their wants, desires, and fears. The second is concerned with the beliefs and attitudes that an agent maintains. Leslie's model has two related but separate mechanisms for dealing with these two classes of events which he calls system-1 and system-2 but which I will refer to as ToMM-1 and ToMM-2 after Baron-Cohen (1995).

ToMM-1 is concerned with actional agency; it deals with agents and the goal-directed actions that they produce. For example, if you see a raccoon slowly approaching a pool of water, you infer that the raccoon might be thirsty, that it *wants* to take a drink. The primitive representations of actions such as approach, avoidance, and escape are constructed by ToMM-1. This system of detecting goals and actions begins to emerge at around 6 months of age (Leslie, 1982). The emergence of ToMM-1 is most often characterized by attention to what other agents are looking at because this serves as a very accurate indicator of intent. Leslie leaves open the issue of whether ToMM-1 is innate or acquired.

ToMM-2 is concerned with attitudinal agency; it deals with the representations of beliefs and how mental states can drive behavior relative to a goal. If the raccoon were to approach a pool of kerosene in the same fashion, you might assume that the raccoon *thought* that it was actually a pool of water. This system develops gradually, with the first signs of development beginning between 18 and 24 months of age and completing sometime near 48 months. ToMM-2 employs the M-representation, a meta-representation which allows truth properties of a statement to be based on mental states rather than observable stimuli. ToMM-2 is a required system for understanding that others hold beliefs that differ from our own knowledge or from the observable world, for understanding different perceptual perspectives, and for understanding pretense and pretending.

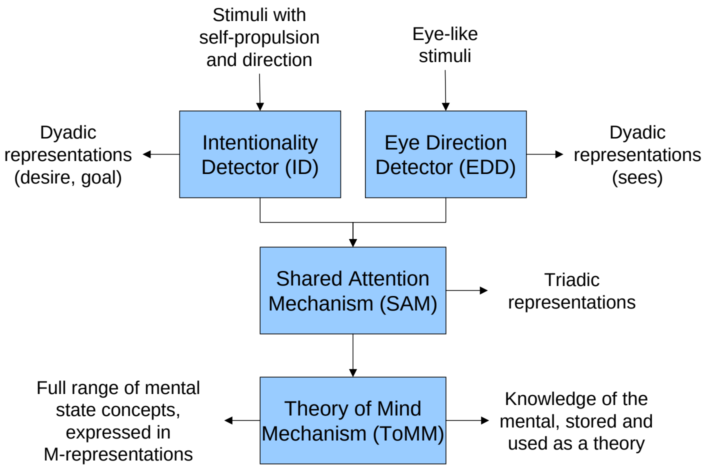

Stimuli with self-propulsion and direction
Eye-like stimuli
Dyadic representations (desire, goal)
Intentionality Detector (ID)
Eye Direction Detector (EDD)
Dyadic representations (sees)
Shared Attention Mechanism (SAM)
Triadic representations
Full range of mental state concepts, expressed in M-representations
Theory of Mind Mechanism (ToMM)
Knowledge of the mental, stored and used as a theory

Figure 3-4: Block diagram of Baron-Cohen's model of the development of theory of mind. See text for description. Adapted from Baron-Cohen (1995).

3.3 Baron-Cohen's Model

While Leslie's model has a clean conceptual division of the external world into three spheres of causality, Baron-Cohen's model is more easily grounded in perceptual processes. Baron-Cohen's model assumes two forms of perceptual information are available as input. The first percept describes all stimuli in the visual, auditory, and tactile perceptual spheres that have self-propelled motion. The second percept describes all visual stimuli that have eye-like shapes. Baron-Cohen proposes that the set of precursors to a theory of mind, which he calls the "mindreading system," can be decomposed into four distinct modules.

The first module interprets self-propelled motion of stimuli in terms of the primitive volitional mental states of goal and desire. This module, called the intentionality detector (ID), produces dyadic representations that describe the basic movements of approach and avoidance. For example, ID can produce representations such as "he wants the food" or "she wants to go over there". This module only operates on stimuli that have self-propelled motion, and thus pass a criterion for distinguishing stimuli that are potentially animate (agents) from those that are not (objects). Baron-Cohen speculates that ID is a part of the infant's innate endowment.

The second module processes visual stimuli that are eye-like to determine the direction of gaze. This module, called the eye direction detector (EDD), has three basic functions. First, it detects the presence of eye-like stimuli in the visual field. Human infants have a preference to look at human faces, and spend more time gazing at the eyes than at other parts of the face. Second, EDD computes whether the eyes are looking at it or at something else. Baron-Cohen proposes that having someone else make eye contact is a natural psychological releaser that produces pleasure in human infants (but may produce more negative arousal in other animals). Third, EDD interprets gaze direction as a perceptual state, that is, EDD codes dyadic representational states of the form “agent sees me” and “agent looking-at not-me”.

The third module, the shared attention mechanism (SAM), takes the dyadic representations from ID and EDD and produces triadic representations of the form “John sees (I see the girl)”. Embedded within this representation is a specification that the external agent and the self are both attending to the same perceptual object or event. This shared attentional state results from an embedding of one dyadic representation within another. SAM additionally can make the output of ID available to EDD, allowing the interpretation of eye direction as a goal state. By allowing the agent to interpret the gaze of others as intentions, SAM provides a mechanism for creating nested representations of the form “John sees (I want the toy)”.

The last module, the theory of mind mechanism (ToMM), provides a way of representing epistemic mental states in other agents and a mechanism for tying together our knowledge of mental states into a coherent whole as a usable theory. ToMM first allows the construction of representations of the form “John believes (it is raining)”. ToMM allows the suspension of the normal truth relations of propositions (referential opacity), which provides a means for representing knowledge states that are neither necessarily true nor consistent with the knowledge of the organism, such as “John thinks (Elvis is alive)”. Baron-Cohen proposes that the triadic representations of SAM are converted through experience into the M-representations of ToMM.

Baron-Cohen (1995) reviews evidence that match the developmental progression of skills observed in infants to the modular decomposition he has proposed. For normal children, ID and the basic functions of EDD are available to infants in the first 9 months of life. SAM develops between 9 and 18 months, and ToMM develops from 18 months to 48 months. However, the most attractive aspects of this model are the ways in which it has been applied both to the abnormal development of social skills in autism and to the social capabilities of non-human primates and other vertebrates.

Baron-Cohen has proposed that the range of deficiencies in autism can be characterized by his model. In all cases, EDD and ID are present. In some cases of autism, SAM and ToMM are impaired, while in others only ToMM is impaired. This can be contrasted with other developmental disorders (such as Down’s Syndrome) or specific linguistic disorders in which evidence of all four modules can be seen.

Furthermore, Baron-Cohen attempts to provide an evolutionary description of these modules by identifying partial abilities in other primates and vertebrates. This phylogenetic description ranges from the abilities of hog-nosed snakes to detect direct eye contact to the sensitivities of chimpanzees to intentional acts. Roughly speaking, the abilities of EDD seem to be the most basic and can be found in part in snakes, avians, and most other vertebrates as a sensitivity to predators (or prey) looking at the animal. ID seems to be present in many primates, but the capabilities of SAM seem to be present only partially in the great apes. The evidence on ToMM is less clear, but it appears that no other primates readily infer mental states of belief and knowledge.

3.4 Implications for Humanoid Robots

A robotic system that possessed a theory of mind would allow for social interactions between the robot and humans that have previously not been possible. A theory of mind would enable social learning, allowing the robot to learn from a human instructor using the same natural social cues that people effortlessly use with each other. No specialized training of the observer would be necessary. The robot would also be capable of expressing its internal state (desires, goals, etc.) in a way that would be naturally interpreted by anyone. Further, a robot that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly. The construction of these systems may also provide a new tool for investigating the predictive power and validity of the human and animal models that serve as the basis. An implemented model can be tested in ways that are not possible to test on humans, using alternate developmental conditions, alternate experiences, and alternate educational and intervention approaches.

The difficulty, of course, is that even the initial components of these models require the coordination of a large number of perceptual, sensorimotor, attentional, and cognitive processes. This section will outline the advantages and disadvantages of Leslie's model and Baron-Cohen's model with respect to implementation. The following section will describe a hybrid architecture that links components of both models with a grounded perceptual and behavioral system.

The most exciting aspect of these models from an engineering perspective is that they attempt to describe the perceptual and motor skills that serve as precursors to the more complex theory of mind capabilities. These decompositions serve as an inspiration and a guideline for building robotic systems that can engage in complex social interactions; they provide a much-needed division of a rather ambiguous ability into a set of observable, testable predictions about behavior. While it cannot be claimed with certainty that following the outlines that these models provide will produce a robot that has the same abilities, the evolutionary and developmental evidence for this skill decomposition does give us hope that these abilities are critical elements of the larger goal. Additionally, the grounding of high-level perceptual abilities to observable sensory and motor capabilities provides an evaluation mechanism for measuring the amount of progress that is being made. Robotic implementations of these systems can be evaluated using the same behavioral and observational metrics that are used to assess the presence or absence of that same skill in children. This decomposition provides a sequence of way-points of testable behavioral skills that can be used to quantitatively measure the progress of a robotic implementation.

Perhaps more importantly, the theory of mind models are interesting from a theoretical standpoint in that they serve as a bridge between skills that are often thought to be high-level cognitive phenomena and low-level skills that are strongly perceptual processes. This link allows for a bottom-up engineering approach to begin to address questions about high-level cognitive tasks by showing how these tasks can be grounded into perceptual and motor capabilities. While this connection may seem obvious given the psychological data, it is often difficult in fields (including robotics) that are driven primarily by bottom-up design to see how these low-level abilities might someday scale to more complex questions. Similarly, in fields (including much of classical artificial intelligence) where top-down design is the status quo, it is difficult to bind abstract reasoning to realistic sensory data. Bottom-up design tends to result in systems that are robust and practical, but that in many ways fail to construct interesting and complex behavior. Top-down design will often result in systems that are elegant abstractions, but that have little hope of being usable in a real system. These models of theory of mind provide the insight to construct a system that is truly grounded in the real-world sensory and motor behaviors but that also can begin to engage some interesting high-level cognitive questions.

From a robotics standpoint, the most salient differences between the two models are the ways in which they divide perceptual tasks. Leslie cleanly divides the perceptual world into animate and inanimate spheres and allows for further processing to occur specifically to each type of stimulus. Baron-Cohen does not divide the perceptual world quite so cleanly but does provide more detail on limiting the specific perceptual inputs that each module requires. In practice, both models require remarkably similar perceptual systems (which is not surprising, since the behavioral data is not under debate). However, each perspective is useful in its own way in building a robotic implementation. At one level, the robot must distinguish between object stimuli that are to be interpreted according to physical laws and agent stimuli that are to be interpreted according to psychological laws. However, the specifications that Baron-Cohen provides will be necessary for building visual routines that have limited scope.

The high-level abstract representations postulated by each model also have implications for robotics. Leslie's model has a very elegant decomposition into three distinct areas of influence, but the interactions between these levels are not well specified. Connections between modules in Baron-Cohen's model are better specified, but they are still less than ideal for a robotics implementation. Additionally, issues on how stimuli are to be divided between the competencies of different modules must be resolved for both models.

3.5 An Embodied Theory of Mind

Drawing from both Baron-Cohen's model and Leslie's model, we propose a hybrid architecture called the *embodied theory of mind*. This model connects modules similar to Leslie's ToBY and Baron-Cohen's EDD, ID, and SAM together with real perceptual processes and with links to physical behaviors. Because both Baron-Cohen and Leslie seek to explain the same underlying data, there is a great deal of overlap in the two representational systems. Leslie's ToMM-1 and ToMM-2 system overlap with the abilities of Baron-Cohen's EDD, ID, SAM, and ToMM modules. However, the emphasis that Leslie places on the theory of body module (ToBY) appears only as an input assumption to Baron-Cohen's model. The embodied theory of mind exploits these overlaps and extends the current models to behavior selection, attention, and

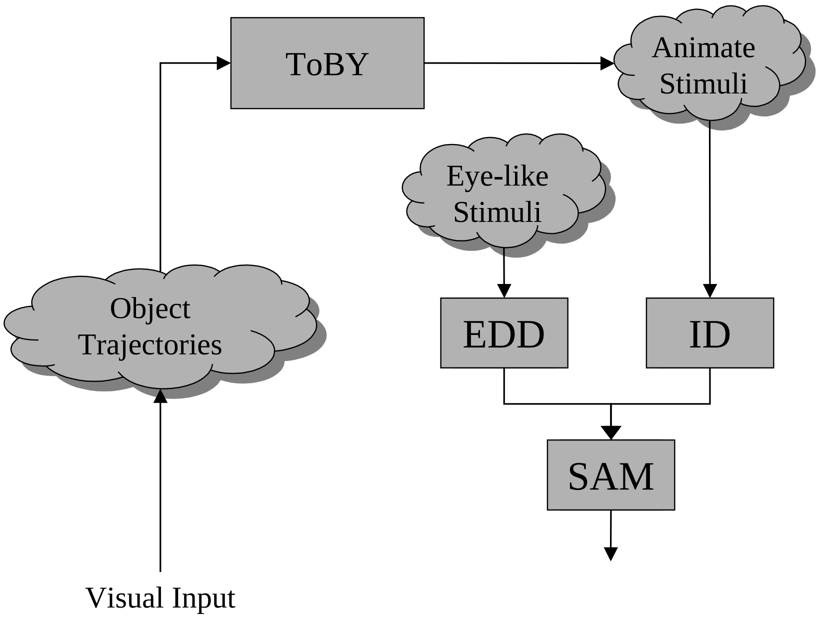


Figure 3-5: The initial stages for linking the Baron-Cohen model and the Leslie model. The primary insight is that the theory of body (ToBY) can serve as a classifier for distinguishing animate from inanimate stimuli.

more complex behavioral forms.

The primary insight in linking the two existing models together is that the theory of body module can act as a classifier for distinguishing self-propelled stimuli. The physical causal laws that ToBY encapsulates are really descriptions of how inanimate objects move through the world. ToBY can be transformed into a classifier by making the assumption that objects that are inanimate must obey these physical laws while objects that are animate will often break them. With this insight, we can begin to sketch out the connections between these modules (see figure 3-5). Visual input will be processed to form motion trajectories, similar to the trajectories observed in Michotte's experiments. These visual trajectories will then be analyzed by a set of naive physical laws in the theory of body module (ToBY). Objects that obey the laws of mechanical causality will be considered to be inanimate, while those that break mechanical causality laws will be classified as animate. Baron-Cohen's model requires two types of input stimuli: objects with self-propelled motion and face-like objects. Animate stimuli trajectories serve directly as the input to Baron-Cohen's intentionality detector (ID). These animate trajectories will also then be processed by additional levels of image processing to find locations that contain faces. These face locations will then be the input to the eye direction detector module (EDD), which then feeds directly to the shared attention mechanism (SAM).

Connecting this rough outline to real perceptual systems and real motor response systems involves slightly more detail but still follows the same general principles. Figure 3-6 shows the overview of the system architecture that will be the subject of

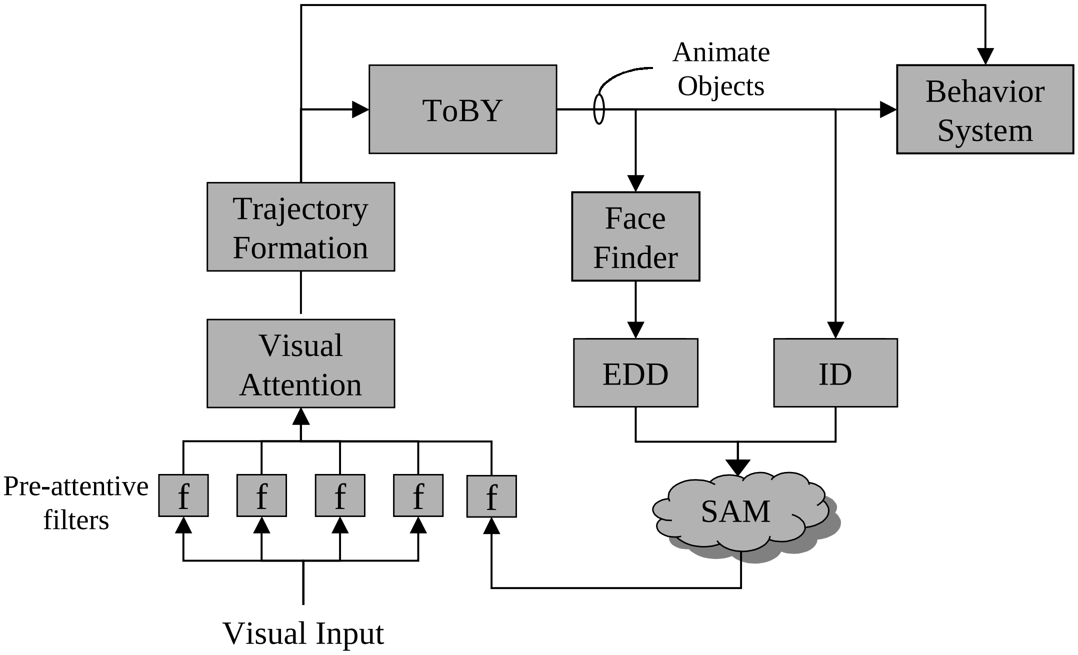


Figure 3-6: Overview of the hybrid theory of mind model.

chapters 6-12. Raw visual input will be processed by a number of low-level feature detectors (such as color, motion, and skin tone) which pre-attentively pick out areas of interest. These low-level filters will be combined with high-level task constraints and a habituation mechanism to select the most salient object in the scene. The attention system performs this selection and then directs limited computational and motor resources to the object of interest (chapter 6). Multiple objects of interest will be produced, and the trajectories of these objects will be tracked through time. These trajectories will serve as the input to the theory of body mechanism, which will employ an agent-based architecture to model the collective knowledge of many simple rules of naive physics (chapter 7). Any objects that violate the naive physical laws will be declared animate and will be subject to further processing by the initial modules of Baron-Cohen's model. Animate stimuli will be processed by a multi-stage face detection system. Any faces in the scene will attract the attention of the robot, which will then use a sequence of post-attentive processing steps to determine the orientation of the individual (chapter 8). These perceptual systems will directly drive behaviors including head orientation, gaze direction, and pointing gestures. In addition, a simple social learning system will be implemented to demonstrate the effects of these social cues on imitative learning (chapter 9). Animate trajectories will also be processed by a simple intentionality detector that picks out relationships between animate objects and other objects based on a simple representation of approach and avoidance (chapter 11). These two representations will trigger shared attention behaviors by applying an additional measurement of object saliency based on the attentional and intentional state of the observed individual (chapter 10). Finally, the extensions of this system toward building a richer set of theory of mind abilities and a more robust representational architecture are discussed in chapter 12.

Before discussing the details of this implementation, chapter 4 describes the three robots that were constructed in part to support this research program. Chapter 5 describes some of the motor and skill learning required to support this implementation.

Chapter 4

Robot Platforms

> *The grand challenge that we wish to take up is to make the quantum leap from experimenting with mobile robot systems to an almost humanoid integrated head system with saccading foveated vision, facilities for sound processing and sound production, and two compliant, dextrous manipulators.* – Brooks & Stein (1994, p. 9)

To address many of the issues raised in chapters 1 and 2, the Humanoid Robotics Group at the MIT Artificial Intelligence Laboratory has been constructing robots that have human-like forms and abilities. To allow humans to engage these robots in natural social interactions, the robots have been designed to mimic the sensory and motor capabilities of the human system. The robots should be able to detect stimuli that humans find relevant, should be able to respond to stimuli in a human-like manner, and should have a roughly anthropomorphic appearance.

This chapter details the design decisions necessary to balance the need for human-like capabilities with the reality of relying on current technology and with design constraints such as reliability, cost, and responsiveness. The primary robotic platform for this work is an upper-torso humanoid robot called Cog. In addition to the work presented here, Cog has been used as an experimental platform for investigations of auditory localization (Irie, 1995), rhythmic arm movements that exploit natural dynamics (Williamson, 1999), learning functional mappings between sensorimotor systems (Marjanović, 1995), and a bio-chemical model of muscle fatigue (Adams, 2000). Cog is also currently an active research platform for work on learning ego-motion relations using sensorimotor correlations (Marjanović, 2001) and for a gestural language for a humanoid robot (Edsinger, 2001).

This chapter will also provide a basic description of two other robots, Kismet and Lazlo, that have been used for portions of this work. Both Kismet and Lazlo are active vision systems that were originally designed as copies of Cog's head, but which have both undergone considerable augmentation. Kismet has been given a set of expressive facial features in order to study the interaction dynamics of the adult-infant relationship (Breazeal, 2000). Lazlo has been given a more anthropomorphic appearance than either Cog or Kismet in order to explore the aesthetic and design

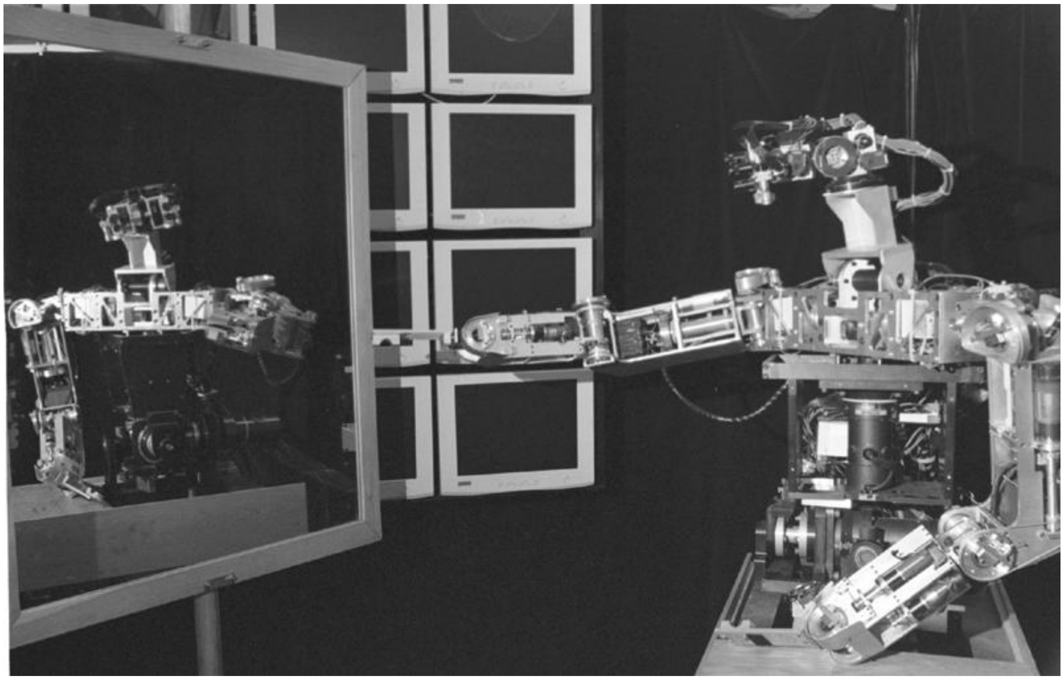


Figure 4-1: Cog, an upper-torso humanoid robot reaching toward an interesting visual stimulus (which happens to be itself in a mirror). The hardware platform has evolved considerably over the last few years. Cog has twenty-two degrees of freedom to approximate human movement, and a variety of sensory systems that approximate human senses, including visual, vestibular, auditory, and tactile senses.

issues involved in building systems that evoke a social response (Edsinger et al., 2000).

4.1 Cog, An Upper-Torso Humanoid Robot

The main robotic platform for this work is an upper-torso humanoid robot called Cog (figure 4-1). The hardware design, computational architecture, and software systems for Cog have undergone considerable evolution over the seven years since the inception of the project. There have been three different mechanical designs for the head, two major revisions of the arm systems, one reconstruction of the torso, three major overhauls in the computational system, and uncountable software changes. This section presents a single snapshot of the development of Cog as of May, 2001, along with the design criteria that have been used in the development of the robot. For a historical perspective on the changes in the robotic systems, see Scassellati (1998*a*), Brooks et al. (1999), and Brooks et al. (1998).

4.1.1 Perceptual Systems

While there are many aspects of human perceptual systems that are far too delicate or detailed to replicate with current hardware technology, we have made attempts to mimic many human perceptual abilities. Mimicking aspects of the human perceptual system provides both additional challenges in the engineering of the system and additional possibilities in the use of human perceptual models. There are many sensory technologies that provide abilities that are super-human in their sensitivity or that are not natural to human beings. For example, many gyroscopes available on the market today provide a higher sensitivity than the human ability to represent orientation through vestibular mechanisms. Other sensor technologies, such as infra-red cameras, sonar, and laser range-finders, are commonly used on robotic systems to provide information that is often difficult or impossible to obtain by normal human senses.

While these technologies can aid in many tasks that are difficult for robotic systems, such as navigation for mobile robotic systems, they are often not the best choices given the goals outlined in chapter 2. Building a system using these super-human abilities is a convenient way to solve some difficult problems, but may lead a researcher away from other fundamental and interesting questions. For example, one common method for obtaining the distance to a target is the use of a laser range-finder. While these systems are moderately large, they have been successfully used on a number of mobile robotic systems. Using this sensing technique, it is possible to avoid the complexities of visual disparity detection, vergence, and depth estimation. If the only goal is to obtain depth estimates, the laser range finder will provide a quicker and easier engineering solution. However, by avoiding the difficulties of visual depth detection, opportunities for using this information to solve problems in visual tracking and object segmentation may be missed.

These additional capabilities may also detract from the social interactions between the robot and others. When based upon these super-human capabilities, the robot's behavior may be uninterpretable by a human observer. If the robot responds to signals that the human cannot sense, the human may attribute the behavior to a very different causal explanation, or may find the behavior completely inexplicable. For example, mobile robots are often equipped with infrared cameras, which aid in finding people in a scene. Imagine that the robot is programmed to provide a greeting every time it encounters a new person, that is, every time it encounters a new heat source. After observing the robot greet a number of people, an observer might be surprised to see the robot greeting a dog, a halogen lamp, or even to greet someone through a thin wall. Or, imagine that a human-like robot had been equipped with a camera that allowed it to see in all directions.[1] If this robot were to face one person and yet respond to another person standing behind it, the social cues that the robot would exhibit become confused.

In building Cog, we have chosen to remain faithful to human senses as much as

[1] Cameras such as these are often used in mobile robotics by pointing a normal camera upward toward a conical mirror.

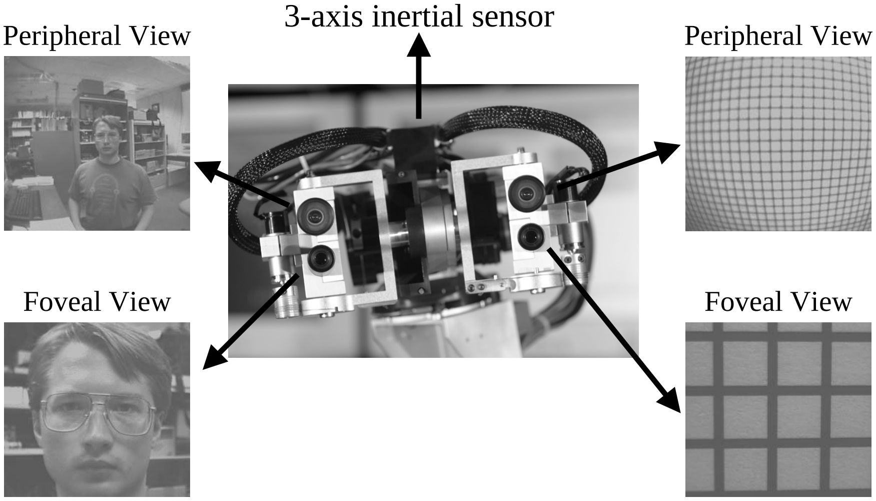


Figure 4-2: Location of the visual and inertial sensors on the robot's head. Each eye contains two color CCD cameras. The upper camera captures a field of view of approximately 120° for peripheral vision. The lower camera captures a smaller field of view ($\approx 20°$) which approximates the human fovea. The images at the left show typical scenes captured from the two cameras. The images at the right were taken by the robot's cameras while pointed toward a piece of graph paper. Notice that while the foveal camera produces a regular grid, the peripheral camera contains a moderate fish-eye distortion. The inertial sensor is mounted above the four degrees of freedom in the neck, and thus moves as the head moves.

possible using commercially available technology. Cog has a variety of sensory systems including visual, vestibular, tactile, and kinesthetic senses.[2] The following sections will provide details on each of the individual sensing systems. Primary emphasis will be placed on the visual system, as the other systems are used only incidentally in the work presented here. Additional information has been published on the auditory system (Irie, 1995) and the kinesthetic system (Williamson, 1999).

Visual System

While current technology does not allow us to exactly mimic all of the properties of the human visual system, there are two properties that we desire: wide field of view and high acuity. Wide field of view is necessary for detecting salient objects in the environment, providing visual context and compensating for ego-motion. High acuity is necessary for tasks like gesture identification, face recognition, and guiding fine motor movements. In a system of limited resources (limited photoreceptors), a balance must be achieved between providing wide field of view and high acuity. In the human retina, this balance results from an unequal distribution of photoreceptors. A

[2] Cog has also had an auditory system at various points in its history, but this was never remounted on the current head design. There are future plans for mounting these microphones once the robot has a face-like shell such as the one currently being developed on Lazlo.

high-acuity central area, called the fovea, is surrounded by a wide periphery of lower acuity. Cog's vision system will also need to balance the need for high acuity with the need for wide peripheral vision (Scassellati, 1998*a*). There are experimental camera systems that provide both peripheral and foveal vision from a single camera, either with a variable density photoreceptor array (van der Spiegel et al., 1989), with space-variant image sampling (Bernardino & Santos-Victor, 1999), or with distortion lenses that magnify the central area (Kuniyoshi et al., 1995). Because these systems are still experimental, factors of cost, reliability, and availability preclude using these options. A simpler alternative is to use two camera systems, one for peripheral vision and one for foveal vision. This alternative allows the use of standard commercial camera systems, which are less expensive, have better reliability, and are more easily available. Using separate foveal and peripheral systems does introduce a registration problem; it is unclear exactly how points in the foveal image correspond to points in the peripheral image. We will address this issue in chapter 8.

The vision system developed for Cog uses four Elmo MN42H remote-head cameras. The remote heads are cylindrical, measuring approximately 17 mm in diameter and 53 mm in length (without connectors), and weighing 25 grams per unit. The upper camera of each eye is fitted with a 3 mm lens that gives Cog a wide peripheral field of view (88.6°(V) × 115.8°(H)). The lower camera is fitted with a 15 mm lens to provide higher acuity in a smaller field of view (18.4°(V) × 24.4°(H)). This creates a fovea region significantly larger than that of the human eye, which is 0.3°, but which is significantly smaller than the peripheral region. Each camera produces an NTSC signal that is digitized by a frame grabber connected to the primary computation system.

Vestibular System

The human vestibular system plays a critical role in the coordination of motor responses, eye movement, posture, and balance. The human vestibular sensory organ consists of the three semi-circular canals, which measure the acceleration of head rotation, and the two otolith organs, which measure linear movements of the head and the orientation of the head relative to gravity.

To mimic the human vestibular system, Cog uses a three-axis inertial sensor from Intersense (www.isense.com). The sensor consists of a single integrated remote package measuring 1.06″ × 1.34″ × 1.2″ and a processing module. The remote sensor is mounted on the robot's head (as shown in figure 4-2) in a position that allows it to move with the robot's head but remain stationary when the eyes are moving (similar to the positioning of our own vestibular organs). The sensor delivers both the angular accelerations in roll, pitch, and yaw and an absolute angular measurement in two dimensions with respect to the gravity vector. The sensor processing module communicates through a standard serial RS-232 interface to the main processing system.

Kinesthetic System

Feedback concerning the state of Cog's motor system is provided by a variety of sensors located at each joint. The eye and head axes use only the simplest form of feedback; each actuator has a single digital encoder which gives position information. The arm joints have the most involved kinesthetic sensing. In addition to all the previous sensors, each of the 12 arm joints also has strain gauges for accurate torque sensing and potentiometers for absolute position feedback.

Other Perceptual Systems

Some previous hardware revisions of Cog incorporated additional sensory capabilities that have either been abandoned or omitted from the current implementation. Other work on Cog has focused on tactile and auditory perception, but these systems have not been integrated into this work.

4.1.2 Motor Systems

To build a system that can engage in interesting social interactions, the motor response of the system must be reliable, flexible, and have a low response latency. The system must also conform to certain safety considerations, allowing people to interact with the robot and to touch the robot. Most importantly, the movement of the robot must be sufficient to evoke a feeling of sociability from the human observer. The robot must move in a natural, biological manner; the movements must be of appropriate speed, timing, and structure. These movements must not only serve to accomplish tasks for the robot, but also to convey social information to the human instructor as to the nature of the action that the robot is performing. In the same way that we naturally interpret the movement of other people and animals, we must be able to interpret the actions of the robot.

Cog's mechanical structure has been designed to approximate both the range of movement and the speed of movement of a human. Cog has a total of twenty-two mechanical degrees of freedom; two arms with six degrees of freedom each, a torso with three degrees of freedom, a neck with three degrees of freedom, and three degrees of freedom in the eyes.

Head and Eyes

Human eye movements can be classified into five categories: three voluntary movements (saccades, smooth pursuit, and vergence) and two involuntary movements (vestibulo-ocular reflex and optokinetic response) (Goldberg et al., 1992). Saccades focus an object on the fovea through an extremely rapid ballistic change in position (up to 900° per second). Smooth pursuit movements maintain the image of a moving object on the fovea at speeds below 100° per second. Vergence movements adjust the eyes for viewing objects at varying depth. While the recovery of absolute depth may not be strictly necessary, relative disparity between objects are critical for tasks such

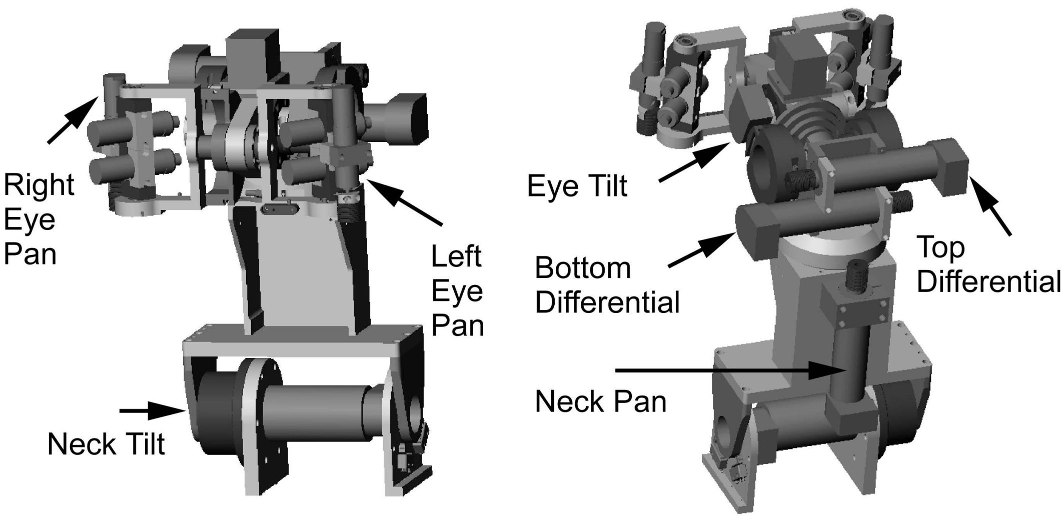


Figure 4-3: The seven mechanical degrees of freedom in Cog's head. The movement of the head has been designed to match the range and speed of human head movement, even though the current outer structure of the head is not a representation of the shape of a human head.

as accurate hand-eye coordination, figure-ground discrimination, and collision detection. The vestibulo-ocular reflex and the optokinetic response cooperate to stabilize the eyes when the head moves.

The goal of mimicking human eye movements generates a number of requirements for the mechanical system. Saccadic movements provide a strong constraint on the design of this system because of the high velocities necessary. To obtain high velocities, the system must be lightweight, compact, and efficient. Smooth tracking motions require high accuracy from the motor control system, and a computational system capable of real-time image processing. Vergence requires a binocular system with independent vertical axis of rotation for each eye. The vestibulo-ocular reflex requires low-latency responses and high accuracy movements, but these requirements are met by any system capable of smooth pursuit. The optokinetic response places the least demanding requirements on this system; it requires only basic image processing techniques and slow compensatory movements.

The active vision system has three degrees of freedom consisting of two active "eyes". Each eye can independently rotate about a vertical axis (pan), and the two eyes share a horizontal axis (tilt). This allows for both conjunctive eye movements, that is, movements in which both eyes move in similar ways in both dimensions, and disjunctive eye movements, in which the two eyes verge in toward the midline or away from the midline. Human eyes have one additional degree of freedom; they can rotate slightly about the direction of gaze. You can observe this rotation as you tilt your head from shoulder to shoulder. This additional degree of freedom is not implemented in our robotic system due to mechanical constraints. To approximate the range of motion of human eyes, mechanical stops were included on each eye to permit a 120°

pan rotation and a 60° tilt rotation. On average, the human eye performs 3 to 4 full range saccades per second (Goldberg et al., 1992). Given this goal, Cog's eye motor system is designed to perform three 120° pan saccades per second and three 60° tilt saccades per second (with 200 milliseconds of stability between saccades). This specification corresponds to angular accelerations of 1309 radians/s^2 and 655 radians/s^2 for pan and tilt.

Cog also has four degrees of freedom in the neck. The *neck tilt* axis brings the whole head forward toward the chest with an axis of rotation near the level of the shoulders. This axis allows for movements that tip the entire head forward and are equivalent to the movement of the upper vertebrae of the spinal column. The *neck pan* axis rotates the entire head about a vertical axis of rotation, allowing the head to look to the left or the right. Finally, a differentially driven set of motors allows for two additional degrees of freedom: the *head roll* which tips the head around an axis of rotation that passes from the center of the head straight out through where the nose would be, and the *head tilt* which nods the head (but not the neck) toward the chest. The *head tilt* axis of rotation can be seen as a line connecting where the robot's ears would be, allowing movements along this axis to be seen as a "yes" nod of the robot's head. These axes of the neck allow the robot to adopt a number of expressive movements including yes/no head nods (movements of the head tilt and neck pan axes respectively), withdrawals and approaches (counter-rotations of the neck tilt and head tilt axes), and looks of curiosity (an approach movement combined with a head roll).

Human observers readily interpret the movement of the head and eyes of the robot as representative of the attentional state and level of commitment of the robot toward a particular object or stimulus (Breazeal et al., 2000*b*). Movements of the eyes alone are easily interpreted as representing the attentional state of the robot. Movements of the eyes followed by an orientation movement of the neck is seen as both an indication of the robot's focus of attention but also as a greater level of interest on the part of the robot.

Arms

Each of Cog's arms is loosely based on the dimensions of a human arm with six degrees of freedom, each powered by a DC electric motor through a series spring (a series elastic actuator, see Pratt & Williamson, 1995). The spring provides accurate torque feedback at each joint and protects the motor gearbox from shock loads. A low gain position-control loop is implemented so that each joint acts as if it were a virtual spring with variable stiffness, damping and equilibrium position. These spring parameters can be changed, both to move the arm and to alter its dynamic behavior. Motion of the arm is achieved by changing the equilibrium positions of the joints, not by commanding the joint angles directly. There is considerable biological evidence for this spring-like property of arms (Zajac, 1989; Cannon & Zahalak, 1982; MacKay et al., 1986).

The spring-like property gives the arm a sensible "natural" behavior: if it is disturbed, or hits an obstacle, the arm simply deflects out of the way. The disturbance is absorbed by the compliant characteristics of the system and needs no explicit sensing or computation. The system also has a low frequency characteristic (large masses and soft springs) which allows for smooth arm motion at a slower command rate. This allows more time for computation and makes possible the use of control systems with substantial delay (a condition akin to biological systems). The spring-like behavior also guarantees a stable system if the joint set-points are fed forward to the arm.

Torso

Cog's torso has three degrees of freedom: the waist bends side-to-side and front-to-back, and the "spine" can twist allowing the arms and head to rotate to the left and right. Each of the degrees of freedom in the torso has recently been modified to use force control based upon feedback from load cells in each of the joints. These load cells provide a torque signal for the amount of force being applied on each joint. One current research project is addressing using the torso and arms in a coordinated fashion (Edsinger, 2001). The mechanical limits and movement of the torso have been calibrated to allow for human-like movements without a loss of safety.

4.1.3 Common Computational Substrate

The computational control for Cog has changed radically over the course of the project. Each revision of the primary computational architecture has focused on providing real-time response for complex signal processing applications with minimal latencies.

The current computational core is a network of off-the-shelf industrial PC computers. There are currently 24 processors, ranging in speed from 200 to 800 MHz, but the network is expandable to 256 processing nodes. Processors are interconnected by a 100 Mbps ethernet with a 1 Gigahertz networking switch, as well as with point-to-point ethernet connections between specific processors. Each processor runs the QNX real-time operating system (www.qnx.com), a commercial product that allows for real-time scheduling in a Linux-like environment. QNX provides transparent and fault-tolerant interprocess communication over the network. The robot is connected to the computational core through commercial video digitization boards (Imagination PCX-200 frame grabbers), through dedicated analog-to-digital conversion boards (from Universal Electronics Industries, www.uei.com), and through commercial motor control boards (from Motion Engineering, www.motioneng.com).

4.2 Kismet, A Robot for Expressive Interaction

Kismet is an active vision head augmented with expressive facial features (see figure 4-4). Kismet is designed to receive and send human-like social cues to a caregiver, who can regulate its environment and shape its experiences as a parent would for a child

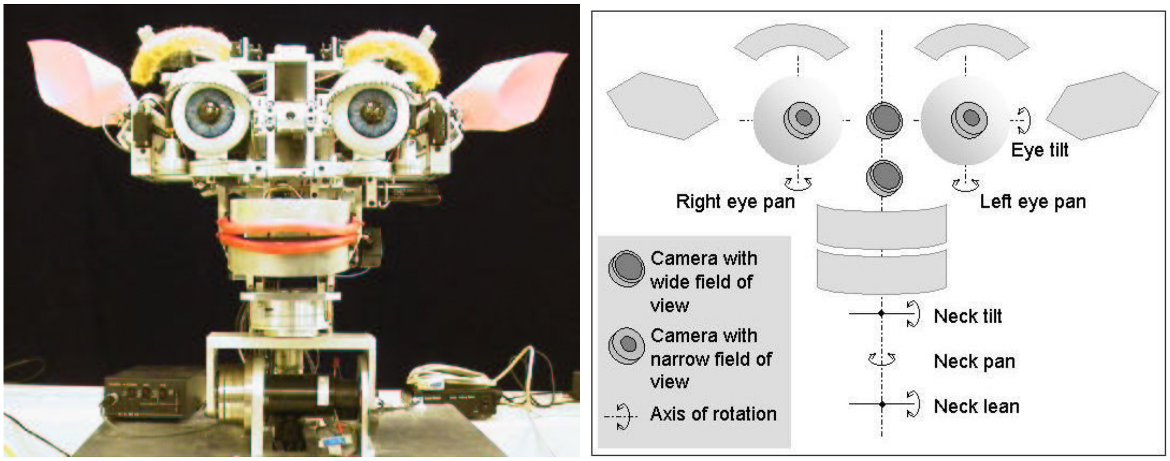
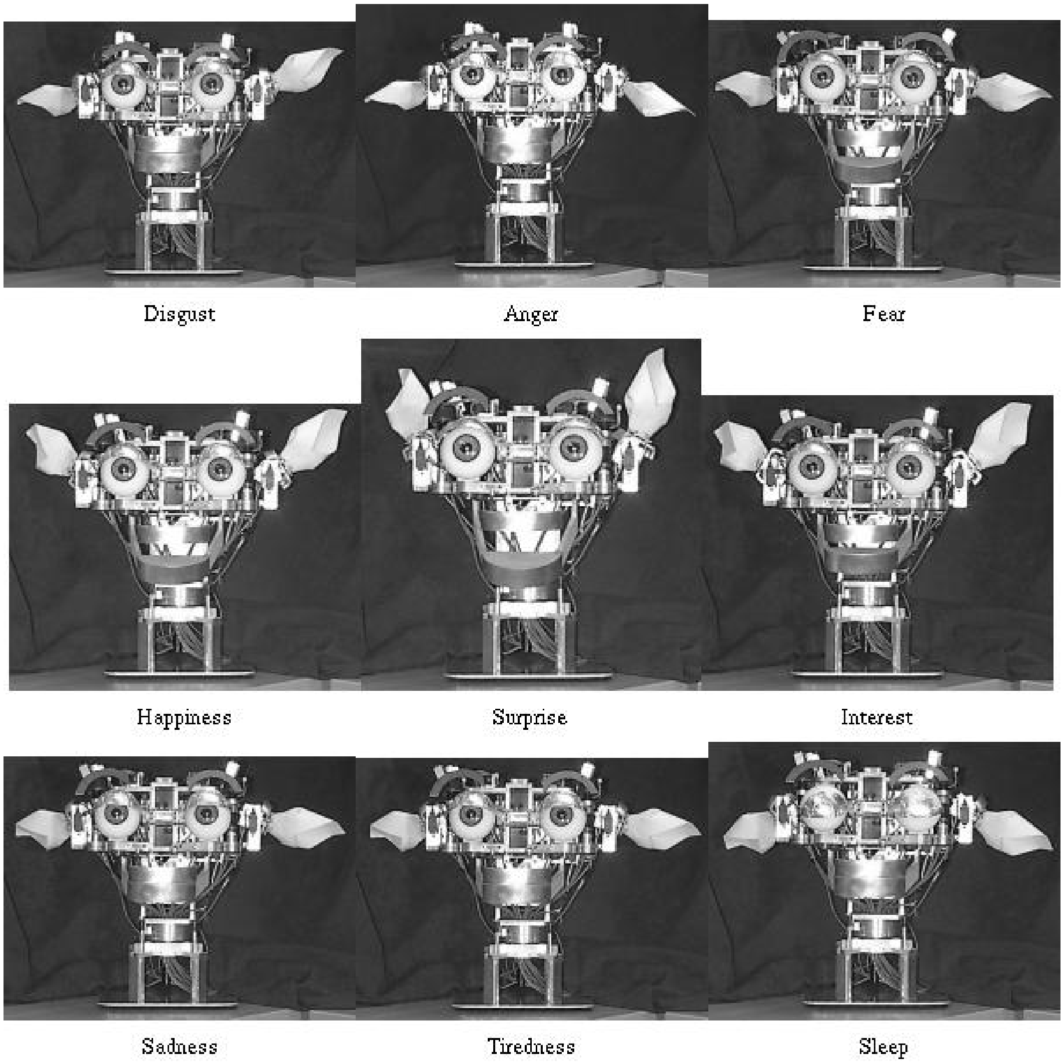


Figure 4-4: Kismet has a large set of expressive features – eyelids, eyebrows, ears, jaw, lips, neck and eye orientation. The schematic on the right shows the degrees of freedom relevant to visual perception (omitting the eyelids!). The eyes can turn independently along the horizontal (pan), but turn together along the vertical (tilt). The neck can turn the whole head horizontally and vertically, and can also crane forward. Two cameras with narrow "foveal" fields of view rotate with the eyes. Two central cameras with wide fields of view rotate with the neck. These cameras are unaffected by the orientation of the eyes. Line drawing courtesy of Paul Fitzpatrick.

(Breazeal, 2000). Kismet has three degrees of freedom to control gaze direction, three degrees of freedom to control its neck, and fifteen degrees of freedom in other expressive components of the face, including eyebrows (each with two degrees of freedom: lift and arch), ears (each with two degrees of freedom: lift and rotate), eyelids (each with one degree of freedom: open/close), a mouth (with one degree of freedom: open/close), and lips which can curl at each of the four corners. The robot is able to show expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise (shown in figure 4-5) which are easily interpreted by an untrained human observer.

To perceive its caregiver, Kismet uses a microphone, worn by the caregiver, and four color CCD cameras. The visual system on Kismet differs from Cog's in the configuration and type of cameras. Kismet has two single-board CCD cameras, one behind each eye, that have a narrow field of view slightly larger than the foveal cameras on Cog. Between the eyes, there are two unobtrusive central cameras fixed with respect to the head, each with a wider field of view but correspondingly lower acuity. This configuration leads to a less anthropomorphic visual sensing system, but has the benefit that certain visual tasks become simpler to implement. For example, smooth-pursuit tracking of a visual stimulus (that is, moving the eyes to maintain the object within the center of the field of view) becomes simpler when the two cameras along the robot's midline are used. Because these cameras do not move with the eyes, the visual processing required for tracking need not compensate for camera motion, leading to a simpler tracking algorithm.

The computational system for Kismet is considerably more heterogeneous than



Disgust Anger Fear

Happiness Surprise Interest

Sadness Tiredness Sleep

Figure 4-5: Static extremes of Kismet's facial expressions. During operation, the 15 degrees of freedom for the ears, eyebrows, mouth, and eyelids vary continuously with the current emotional state of the robot.

Cog's, although the components used for vision are nearly identical. Kismet's vision system is implemented on a network of nine 400 MHz commercial PCs running the QNX real-time operating system. Kismet also has a motivational system which runs on a collection of four Motorola 68332 processors. Machines running Windows NT and Linux are also networked for speech generation and recognition respectively.

4.3 Lazlo, A Visual Development Platform

A third robot, called Lazlo, was constructed to provide a second development platform and to allow additional researchers to contribute to the visual processing system (see figure 4-6). Until the most recent revision, Lazlo was an exact copy of the hardware

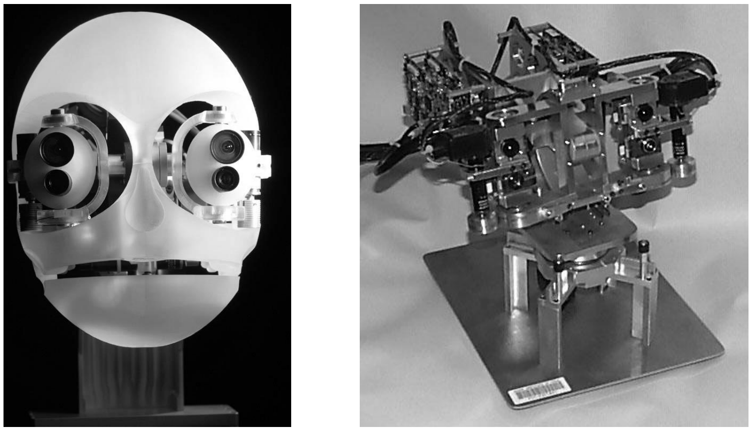


Figure 4-6: In addition to Cog and Kismet, a third development platform called Lazlo was constructed. The earliest version of Lazlo (shown at right) was used primarily as a visual skill development platform. The most current revision (left),has been modified to have a more anthropomorphic appearance.

architecture that was used for Cog's head. The most recent revision is a copy of the understructure of Cog's head and neck, but with additional mount points and a new "face". These additions were intended to allow research on the aesthetic qualities that enable more natural social interactions and the development of social contracts. The underlying perceptual and computational systems for Lazlo are identical to those on Cog. Additional details on the original development platform design can be found in Scassellati (1998*a*), and the most recent revision is described in Edsinger et al. (2000).

4.4 Summary

Three robots were designed and constructed by the Humanoid Robotics Group at MIT to support natural social interactions with people. Cog has the most sensory and manipulation capabilities, and has been the centerpiece of many related projects. With a human-like form, human-like sensory systems, and human-like movements, Cog is a unique platform for investigating how people use and respond to social cues. Kismet is an active vision system that has been modified to have expressive capabilities that help to engage people and to regulate its interactions with the world. Lazlo is a development platform for visual processing routines and for investigating the aesthetics of social interaction. All three robots were used to support the work described in the following chapters.

Chapter 5

Matching Human Movements

> *Because people have a strong positive bias toward social relationships and predictable environments, the more a media technology is consistent with social and physical rules, the more enjoyable the technology will be to use. Conforming to human expectations means that there is instant expertise just because we're human...* – Reeves & Nass (1996, p. 8–9)

In addition to providing an anthropomorphic appearance, we want our robots to move in human-like ways. By matching human movement behaviors, the robot's behavior will be easily understood by a human observer because it is analogous to the behavior of a human in similar circumstances. For example, when an anthropomorphic robot moves its eyes and neck to orient toward an object, an observer can effortlessly conclude that the robot has become interested in that object. By creating behaviors that match human behaviors, the robot can more easily be seen as fitting into the expected social norms. There are other advantages to modeling our implementation after the human motor system. There is a wealth of data and proposed models for explaining human and animal motor responses. This data provides both a standard with which to judge our implementations and often a well-recognized set of evaluation metrics for measuring the progress of the robot's motor behaviors.

This chapter reviews a set of behavioral responses that have been developed for Cog, Kismet, and Lazlo, so that the later chapters on foundational skills for a theory of mind can be evaluated in context. The reader is referred to Scassellati (1999*a*); Brooks et al. (1999); Scassellati (1998*a*); Breazeal et al. (2000*a*); Marjanović et al. (1996), and the other references throughout this chapter for a more extensive review of the appropriate motor responses.

Whenever possible, we have attempted to build adaptive systems that learn to perform sensorimotor skills rather than using explicitly specified kinematic models. This constraint allows the same software to be usable across all of the robotic platforms, even though the kinematics of each will differ slightly. This also allows for more robust behavior, as the kinematic and dynamic aspects of any motor system will change gradually over time due to slight adjustments in the system and mechanical wear.

5.1 Eye Movements

As described in section 4.1.2, human eye movements can be classified into five categories: three voluntary movements (saccades, smooth pursuit, and vergence) and two involuntary movements (vestibulo-ocular reflex and opto-kinetic response)(Goldberg et al., 1992). We have implemented mechanical analogs of each of these eye motions (Scassellati, 1999*a*; Marjanović et al., 1996).

5.1.1 Saccades

Saccades are high-speed ballistic motions that focus a salient object on the high-resolution central area of the visual field (the fovea). In humans, saccades are extremely rapid, often up to 900° per second. To enable our machine vision systems to saccade to a target, we require a saccade function $S : (\vec{x}, \vec{e}) \mapsto \Delta\vec{e}$ which produces a change in eye motor position ($\Delta\vec{e}$) given the current eye motor position ($\vec{e}$) and the stimulus location in the image plane ($\vec{x}$). To obtain accurate saccades without requiring an accurate model of the kinematics and optics, a self-supervised learning algorithm estimates the saccade function. This implementation can adapt to the non-linear optical and mechanical properties of the vision system.

Distortion effects from the wide-angle lens create a non-linear mapping between the location of an object in the image plane and the motor commands necessary to foveate that object. One method for compensating for this problem would be to exactly characterize the kinematics and optics of the vision system. However, this technique must be recomputed not only for every instance of the system, but also every time a system's kinematics or optics are modified in even the slightest way. To obtain accurate saccades without requiring an accurate kinematic and optic model, we use a self-supervised learning algorithm to estimate the saccade function.

Marjanović et al. (1996) learned a saccade function for this hardware platform using a 17×17 interpolated lookup table. The map was initialized with a linear set of values obtained from self-calibration. For each learning trial, a visual target was randomly selected. The robot attempted to saccade to that location using the current map estimates. The target was located in the post-saccade image using correlation, and the L_2 offset of the target was used as an error signal to train the map. The system learned to center pixel patches in the peripheral field of view. The system converged to an average of < 1 pixel of error in a 128×128 image per saccade after 2000 trials (1.5 hours). With a trained saccade function S, the system can saccade to any salient stimulus in the image plane. We have used this mapping for saccading to moving targets, bright colors, and salient matches to static image templates.

Saccade map training begins with a linear estimate based on the range of the encoder limits (determined during self-calibration). For each learning trial, we generate a random visual target location (x_t, y_t) within the 128×128 image array and record the normalized image intensities $\bar{I}_t$ in a 13×13 patch around that point. The reduced size of the image array allows us to quickly train a general map, with the possibility for further refinement after the coarse mapping has been trained. Once the random

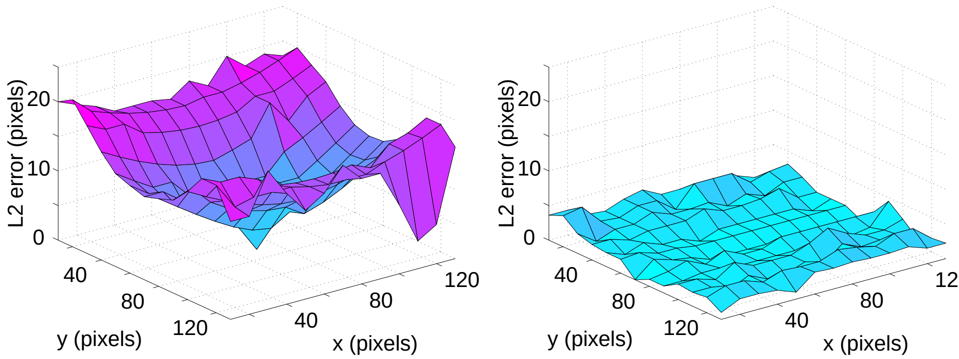


Figure 5-1: L_2 error for saccades to image positions (x,y) after 0 training trials (left) and 2000 training trials (right) using an interpolated lookup table.

target is selected, we issue a saccade motor command using the current map estimate. After the saccade, a new image $\bar{I}_{t+1}$ is acquired. The normalized 13×13 center of the new image is then correlated against the target image. Thus, for offsets x_0 and y_0, we seek to maximize the dot-product of the image vectors:

$$\max_{x_0,y_0} \left(\sum_i \sum_j \bar{I}_t(i,j) \cdot \bar{I}_{t+1}(i + x_0, j + y_0) \right) \tag{5.1}$$

Because each image was normalized by the average luminance, maximizing the dot product of the image vectors is identical to minimizing the angle between the two vectors. This normalization also gives the algorithm a better resistance to changes in background luminance as the camera moves. In our experiments, we only examine offsets x_0 and y_0 in the range of $[-32, 32]$. The offset pair that maximized the expression in equation 5.1, scaled by a constant factor, is used as the error vector for training the saccade map.

Figure 5-1 shows the L_2 error distance for saccades after 0 learning trials and after 2000 trials. After 2000 training trials, an elapsed time of approximately 1.5 hours, training reaches an average L_2 error of less than 1 pixel. As a result of moving objects during subsequent training and the imprecision of the correlation technique, this error level remained constant regardless of continued learning.

We have also used the same training data with different function approximation techniques including neural networks and spline fitting. In each of these cases, the approximated functions have similar error curves and similar times to convergence (Scassellati, 1999*a*).

5.1.2 Smooth Pursuit Tracking

Smooth pursuit movements maintain the image of a moving object on the fovea at speeds below 100° per second. Our current implementation of smooth pursuit tracking acquires a visual target and attempts to maintain the foveation of that target using a cross-correlation metric. Following a saccade, a new target is acquired and installed as the correlation mask by extracting the central $m_h \times m_w$ pixels from the post-saccade image. In subsequent frames, the correlation mask is convolved with each position in a search region of size $s_h \times s_w$ to produce correlation scores $X_{(i,j)}$ where $i \in [1...s_h]$ and $j \in [1...s_w]$. The position with the lowest cross-correlation value is considered to be the center of the new target. A more robust mechanism would use segmentation to delimit the target within the image, but this simple scheme has proved successful for many real-world interactions with the robot.

To ensure that the tracking signal accurately reflects a valid match, three criteria are imposed to maintain consistency. First, the best score must pass a threshold test. This ensures that a sudden, extremely poor match will not cause an erratic eye movement. Second, the quality of the match must exceed a threshold value. The quality of the match is defined as :

$$Q = (\max_{i,j} X_{(i,j)} - \min_{i,j} X_{(i,j)})/(s_h * s_w);$$

Intuitively, this criteria ensures that the best match location is significantly better than the worst possible match, which prevents the system from selecting from among many similar options. The quality also ensures that the system will not move when the correlation mask is a poor match to all of the possible search locations. Third, the average correlation score for all evaluated locations within a single image must also pass a threshold test. This prevents the tracker from wandering randomly when presented with a blank background or with a poor quality mask. When these three criteria are satisfied, the target is declared valid and is used to generate eye movement. If any of these criteria fails, the match is declared invalid. If a consecutive sequence of m matches are declared invalid, the tracker declares that it is lost, which triggers a saccade to acquire a new target.

The vector from the current image center to the center of the valid match is used as a visual error signal, which is then scaled by a constant vector to generate a velocity signal for the eye motors. In practice, for an image of size 128×128, target masks of size $m_h = 8$ by $m_w = 8$ are used with a search area of $m_h = 40$ by $m_w = 40$. This allows for tracking at real-time rates (30 Hz).

5.1.3 Binocular Vergence

Vergence movements adjust the eyes for viewing objects at varying depth. While the recovery of absolute depth may not be strictly necessary, relative disparity between objects is critical for tasks such as accurate hand-eye coordination, figure-ground discrimination, and collision detection. A variety of different computational techniques have been used to provide enough depth information to drive vergence movements (e.g., Rougeaux & Kuniyoshi, 1997; Coombs & Brown, 1993; Yeshurun & Schwartz, 1989)

We have re-implemented the zero-disparity filtering technique used by Coombs & Brown (1993) to drive vergence. We have not yet incorporated this system to consistently drive vergence on Cog, but we have used the system to detect whether two objects exist at the same depth. This information will be used later in chapter 7 to detect potential elastic collisions. Because the moving objects may not be on the depth plane defined by the current vergence angle of the eyes, we have made a slight alteration of the zero-disparity filter. For a given object, the correlation-based tracker is used to locate the same object in the left and right eye images. The difference in position of the target between the two images defines a disparity shift, which can be used to shift the left (or right) image so that the two objects are aligned at the same coordinate location. The zero-disparity filter is then applied to these two shifted images to find other patches in the image at that depth.

5.1.4 Vestibular-Ocular and Opto-Kinetic Reflexes

The vestibulo-ocular reflex and the opto-kinetic nystigmus cooperate to stabilize the eyes when the head moves. The vestibulo-ocular reflex (VOR) stabilizes the eyes during rapid head motions. Acceleration measurements from the semi-circular canals and the otolith organs in the inner ear are integrated to provide a measurement of head velocity, which is used to counter-rotate the eyes and maintain the direction of gaze. The opto-kinetic nystigmus (OKN) compensates for slow, smooth motions by measuring the optic flow of the background on the retina (also known as the visual slip). OKN operates at much lower velocities than VOR (Goldberg et al., 1992). Many researchers have built accurate computational models and simulations of the interplay between these two stabilization mechanisms (Lisberger & Sejnowski, 1992; Panerai & Sandini, 1998).

A simple OKN can be constructed using a rough approximation of the optic flow on the background image. Because OKN needs only to function at relatively slow speeds (5 Hz is sufficient), and because OKN only requires a measurement of optic flow of the entire field, the computational load is manageable. A standard optic flow algorithm (Horn, 1986) calculates the full-field background motion between successive frames, giving a single estimate of camera motion. The optic flow estimate is a displacement vector for the entire scene. Using the saccade map, an estimate of the amount of eye motion required to compensate for the visual displacement can be estimated.

A simple VOR can be constructed by integrating the velocity signal from the inertial system, scaling that signal, and using it to drive the eye motors. This technique works well for transient and rapid head motions, but fails for two reasons. First, because the gyroscope signal must be integrated, the system tends to accumulate drift. Second, the scaling constant must be selected empirically. Both of these deficits can be eliminated by combining VOR with OKN.

Combining VOR with OKN provides a more stable, robust system. The OKN system can be used to train the VOR scale constant. The training routine moves the neck at a constant velocity with the VOR enabled. While the neck is in motion, the OKN monitors the optical slip. If the VOR constant is accurate for short neck motions, then the optical slip should be zero. If the optical slip is non-zero, the VOR constant can be modified in the appropriate direction. This on-line technique can adapt the VOR constant to an appropriate value whenever the robot moves the neck at constant velocity over short distances. The combination of VOR and OKN can also eliminate gradual drift. The OKN will correct not only for slow head motions but also for slow drift from the VOR. We are currently working on implementing models of VOR and OKN coordination to allow both systems to operate simultaneously.

An alternative to the VOR/OKN mechanisms for gaze stabilization is the use of efference copy. In efference copy, motor command signals for neck movement are copied, scaled appropriately, and then sent to counter-rotate the eyes:

$$\Delta\vec{e} = \vec{k} \times \vec{n}$$

The change in eye position ($\Delta\vec{e}$) is the product of the neck position ($\vec{n}$) with a scalar vector ($\vec{k}$). Similar to the training for VOR, the scale vector $\vec{k}$ can be estimated by observing the image slip while a given neck command is executed. In practice, the scale factor between the neck pan axis and the eye pan axes is -1.1, the scale factor between the neck tilt axis and the eye tilt axis is 1.0, and the scale factor between the virtual head tilt axis (a combination of the two differential axes) and the eye tilt axis is -1.45. The sign of the scale factor reflects that the two axes are either wired to move in the same direction (negative scale factors, since this is a mapping that should counter-rotate the eyes) or to move in opposite directions (positive scale factors).

While this mechanism is effective only for self-induced movement, it is more reliable than inertial sensing. The efferent copy signal to the eye motors will arrive with minimal latency. Even a simple control loop for VOR will impose additional delay on the signal. This near-zero latency response can reduce image blur even for a properly tuned VOR/OKN system for movements of the head and neck. However, compensating for movement of the torso imposes an additional difficulty in translating between the force-controlled axes of the torso and the position-controlled axes of the eyes. There is current research on developing a system that learn these relationships (Marjanović, 2001). In practice, we use the efference copy system whenever the neck is moving and a VOR response at all other times. The gains for the VOR system were chosen empirically.

5.2 Coordinated Eye and Neck Movements

Orienting the head and neck along the angle of gaze can maximize the range of the next eye motion while giving the robot a more life-like appearance. Head orientation movements have a very strong social influence. This orientation movement is a strong indicator of social engagement and is easily interpreted by a human observer (Breazeal et al., 2000*b*,*a*).

Once the eyes have foveated a salient stimulus, the neck should move to point the head in the direction of the stimulus while the eyes counter-rotate to maintain fixation on the target (see figure 5-2). To move the neck the appropriate distance, we use a mapping $N : (\vec{n}, \vec{e}) \mapsto \Delta\vec{n}$ to produce a change in neck motor positions ($\Delta\vec{n}$)

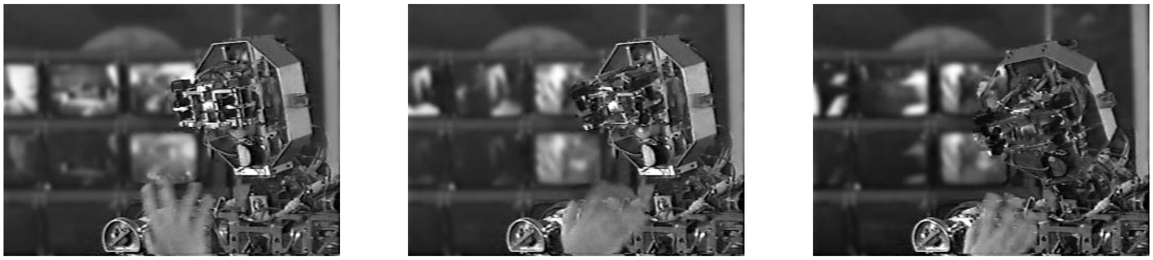


Figure 5-2: Orientation to a salient stimulus. Once a salient stimulus (a moving hand) has been detected, the robot first saccades to that target and then orients the head and neck to that target.

given the current neck position ($\vec{n}$) and the initial eye position ($\vec{e}$). Because the axes of rotation for the eyes are parallel to the appropriate axes of rotation of the head, a simple linear mapping has sufficed: $\Delta\vec{n} = (\frac{1}{k} \times \vec{e} - \vec{n})$ where $\vec{k}$ is the same constant factor used for efference copy.[1]

5.3 Arm Movements

Cog's arms have been designed to support force-control strategies for rhythmic arm movements (Williamson, 1999). The majority of the research on arm control within the Cog project has focused on exploiting the natural dynamics and force feedback to perform rhythmic control movements including turning cranks, swinging pendula, sawing through boards, and playing the drum. However, a few ballistic arm movements have also been studied on this platform. This section will describe two ballistic movements: pointing to a visual target and following a visual trajectory. While this is certainly not a complete list, these two behaviors will be useful for building a system that can mimic human movements.

5.3.1 Pointing to a Visual Target

The ability to point to a visual target appears in infants sometime near 9 months of age (Baron-Cohen, 1995). At this age, an infant will reach towards objects that are of interest, often with the hand open and palm extended toward the object and occasionally opening and closing the hand to indicate interest. This ability is also sometimes referred to as *imperative pointing* to distinguish it from the later-developing action called *declarative pointing* which is used to direct the attention of the parent to an object of interest (Gomez, 1991). This ability is also believed to be a critical part in learning to reach for objects of interest (Diamond, 1990).

There are many ways to approach the problem of enabling a robot to point to a visual target. If we consider the head to be in a fixed position, a purely kinematic

[1] This linear mapping has only been possible with motor-motor mappings and not sensorimotor mappings because of non-linearities in the sensors.

solution is a $R^2 \rightarrow R^4$ sensorimotor mapping problem with no obvious training signal; the position of the target in the visual coordinates (a two-dimensional quantity) must be converted into an arm trajectory for the four degrees of freedom in the arm which are involved in positioning the end effector. While the arm has six mechanical degrees of freedom, the action of pointing is only a four dimensional problem. One mechanical degree of freedom rotates the forearm about its principle axis, and while this alters the orientation of the hand, it does not change the direction of pointing. Further, two degrees of freedom, one at the shoulder and one at the elbow, both produce movement in a plane that is perpendicular to the line connecting the robot's shoulders and thus form a redundant system. However, even this $R^2 \rightarrow R^4$ mapping is still too large a search space to allow for random explorations. Furthermore, it is unclear how to obtain a reliable error signal for pointing. With head movements, the dimensionality of the mapping problem becomes even more complex: $R^6 \rightarrow R^4$.

To simplify the dimensionality problem associated with learning to point and to uncover reliable error signals, we have applied two different aspects of the methodology discussed in chapter 2. The first implementation uses a developmental decomposition of pointing behavior based on the progression of stages that infants pass through in learning to reach. The benefit of this method is that it is completely self-trained, and can be done without human effort. The second implementation uses a set of social constraints to provide the robot with appropriate learning examples. This method requires the assistance of a benevolent instructor to assist the robot in learning to point.

Self-Trained Pointing

Diamond (1990) has shown that infants between five and twelve months of age progress through a number of distinct phases in the development of visually guided reaching. In this progression, infants in later phases consistently demonstrate more sophisticated reaching strategies to retrieve a toy in more challenging scenarios. Using the behavioral decomposition Diamond (1990) observed in infants, Marjanović et al. (1996) implemented a system that learns to point toward a visual target. The implemented system simplifies the dimensionality of the process and allows for the robust recovery of training signals. Given a visual stimulus, typically by a researcher waving an object in front of its cameras, the robot saccades to foveate on the target, and then reaches out its arm toward the target (see figure 5-3). Early reaches are inaccurate, and often in the wrong direction altogether, but after a few hours of practice the accuracy improves drastically.

To reach to a visual target, the robot must learn the mapping from the target's image coordinates $\vec{x} = (x, y)$ to the coordinates of the arm motors $\vec{\alpha} = (\alpha_0 ... \alpha_5)$ (see figure 5-4). To achieve this, the robot first learns to foveate the target using the saccade map $\vec{S} : \vec{x} \rightarrow \vec{e}$ which relates positions in the camera image with the motor commands necessary to foveate the eye at that location. This foveation guarantees that the target is always at the center of the visual field. The reaching movement considers only the 2-D projected position of the target on the image plane without regard for depth. Once the target is foveated, the joint configuration necessary to

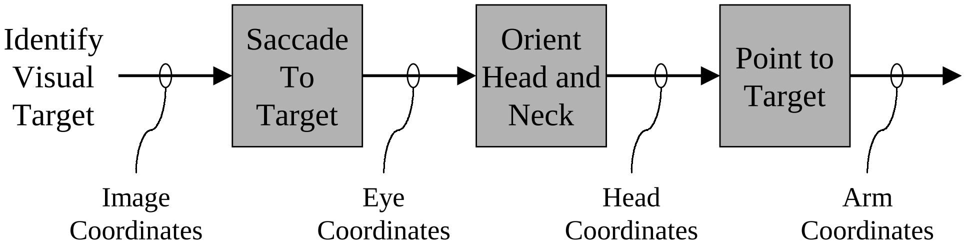


Figure 5-3: A developmental decomposition of reaching behavior. The implemented system considers the neck to be in a fixed position, but the addition of neck movement could easily be accomplished using the orientation behavior already discussed.

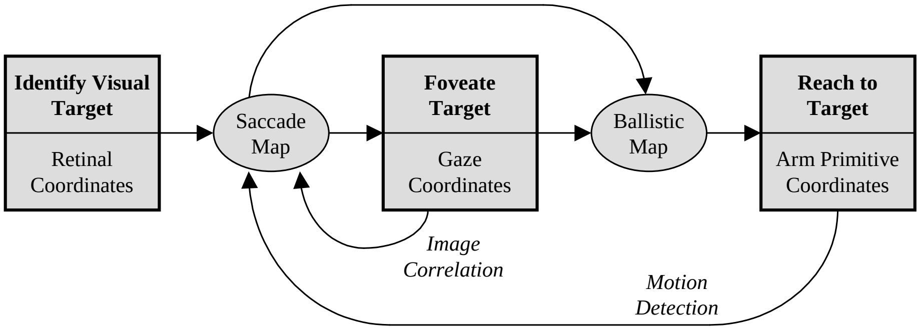


Figure 5-4: Training signals for learning to point to a visual target. Pointing is the product of two sub-skills: foveating a target and generating a ballistic reach from that eye position. Image correlation can be used to train a saccade map which transforms retinal coordinates into gaze coordinates (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.

point to that target is generated from the gaze angle of the eyes using a "ballistic map."

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. Although the arm has four joints active in moving the hand to a particular position in space (the other two control the orientation of the hand), we re-parameterize in such a way that we only control two degrees of freedom for a reach. The position of the outstretched arm is governed by a normalized vector of postural primitives (Mussa-Ivaldi et al., 1985). A primitive is a fixed set of joint angles, corresponding to a static position of the arm, placed at the corners of the workspace. Three such primitives form a basis for the workspace. The joint-space command for the arm is calculated by interpolating the joint-space components between each primitive, weighted by the coefficients of the primitive-space vector. Since the vector in primitive space is normalized, three coefficients give rise to only two degrees of freedom. Hence, a mapping between eye gaze position and arm position, and vice versa, is a simple, non-degenerate $R^2 \rightarrow R^2$ function. This considerably simplifies

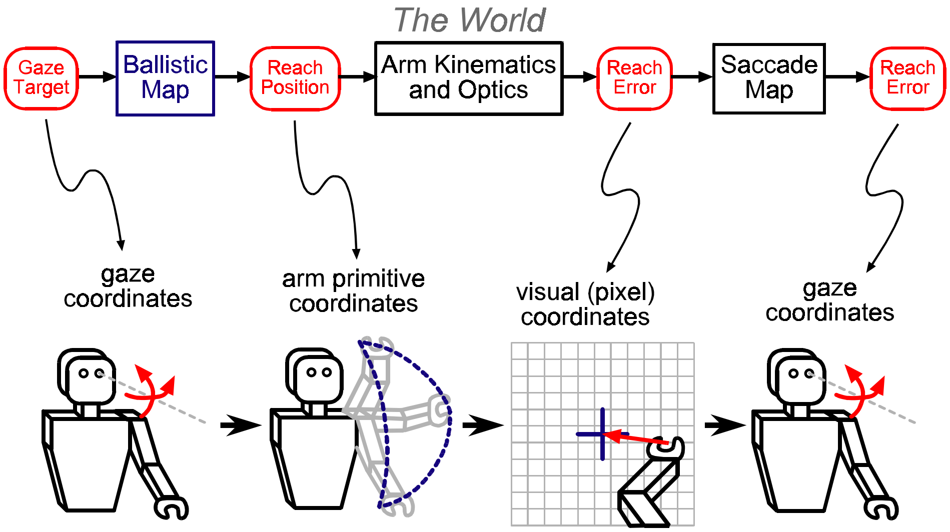


Figure 5-5: Generation of error signals from a single reaching trial. Once a visual target is foveated, the gaze coordinates are transformed into a ballistic reach by the ballistic map. By observing the position of the moving hand, we can obtain a reaching error signal in image coordinates, which can be converted back into gaze coordinates using the saccade map.

learning.

Training the ballistic map is complicated by the inappropriate coordinate space of the error signal. When the arm is extended, the robot waves its hand. This motion is used to locate the end of the arm in the visual field. The distance of the hand from the center of the visual field is the measure of the reach error. However, this error signal is measured in units of pixels, yet the map being trained relates gaze angles to joint positions. The reach error measured by the visual system cannot be directly used to train the ballistic map. However, the saccade map has been trained to relate pixel positions to gaze angles. The saccade map converts the reach error, measured as a pixel offset on the retina, into an offset in the gaze angles of the eyes (as if Cog were *looking* at a different target). In this way, the knowledge gained from learning to foveate a target transforms the ballistic arm error into an error signal that can be used to train the arm directly (see figure 5-5). This re-use allows the learning algorithms to operate continually, in real time, and in an unstructured "real-world" environment without using explicit world coordinates or complex kinematics.

This is still not enough to train the ballistic map. Our error is now in terms of gaze angles, not joint positions (i.e., we know the gaze position that would have foveated the visual target, but not how the arm should move to attain that position). To train the ballistic map, we also need a "forward map," that is, a forward kinematics function which gives the gaze angle of the hand in response to a commanded set of joint positions (Jordan & Rumelhart, 1992). The error in gaze coordinates can be back-propagated through this map, yielding a signal appropriate for training the ballistic map.

The forward map is learned incrementally during every reach: after each reach we know the commanded arm position, as well as the position measured in eye gaze coordinates (even though that was not the target position). For the ballistic map to train properly, the forward map must have the correct signs in its derivative. Hence, training of the forward map begins first, during a "flailing" period in which Cog performs reaches to random arm positions distributed through its workspace.

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). One limitation of this work is that the notion of postural primitives as formulated is very brittle: the primitives are chosen ad-hoc to yield a reasonable workspace. Finding methods to adaptively generate primitives and divide the workspace is a subject of active research.

Socially-Trained Pointing

Another method for simplifying the problem of learning to point is to rely upon social contracts to provide appropriate feedback to the learning algorithm. In this case, we rely upon the presence of a benevolent caregiver to structure the environment in such a way that the learning algorithm is always presented with a partial success. The driving observation for this method is that when children attempt to point to an object, the parent will often respond by providing the child with that object, or by moving an object into the reach of the child. In this way, the parent is always acting to provide the child with a positive example; the location of the presented object and the executed reach constitute a positive example for the ballistic mapping function.

In collaboration with Bryan Adams, a prototype system for socially-trained reaching has been implemented. Four primitive postures were defined for each arm. The first three postures were with the arm near full extension and were used to define a manifold of possible end-points within the robot's range of reaching. The first posture was near the center of the workspace, roughly pointing straight out in front of the robot. The second posture was at roughly the same horizontal position as the first primitive but extended upward toward the limit of the workspace. Similarly, the third posture was at the same vertical position as the first primitive but extended horizontally toward the extreme right of the robot's body. These three postures defined a manifold of extended arm postures. Any point within that surface could be localized as a linear combination of the first posture with some percentage of the displacements toward the second and third postures. The fourth postural primitive was a rest posture in which the robot's arm hung limp along the side of its body. A pointing gesture was defined as a smooth movement from the rest posture to a posture that was a linear combination of the first three postures.

Using this set of postural primitives, the end position of a pointing gesture could be defined by two values. The vertical arm posture A_v was the percentage of the second postural primitive that was to be added to the first postural primitive, and the horizontal arm posture A_h was the percentage of the third postural primitive.

The total end posture of the arm P^* was defined to be:

$$A^* = P_1 + A_v * (P_2 - P_1) + A_h * (P_3 - P_1)$$

where P_i is the i^{th} postural primitive and the scalar values A_v and A_h were allowed to range between 1 and -1. On each learning trial, the robot would point to a random position by selecting random values of A_v and A_h. A human acting as the parent would then provide the robot with a positive learning example by picking up an object near where the robot pointed, moving it to be in alignment with the robot's gesture, and making the object more salient by shaking it, attending to it, and presenting it toward the robot.[2] For each trial then, the robot was presented with a single example of where it would need to reach given a visual position. Approximately 100 trials were recorded in a session lasting slightly more than ten minutes. The data that was collected was processed off-line, although on-line algorithms similar to those used in the developmental decomposition could easily be used. Figure 5-6 shows the resulting visual image coordinates of the most salient object in the scene given the starting arm posture values. Polynomial least-squares fitting revealed a linear fit between the horizontal arm posture and the image column and a quadratic fit between the vertical arm posture and the image row. By inverting these functions, we obtain a reaching function that provides a set of posture ratios given an image location (r, c) as input:

$$A_v = 0.0134 * r^2 - 0.6070 * r + 6.8009 \tag{5.2}$$

$$A_h = 1.4819 * c - 51.2754 \tag{5.3}$$

Examining the basis postures reveals that the horizontal posture does indeed create a nearly-horizontal movement through the image plane that varies linearly, while the vertical posture creates a slightly non-linear projection as the arm reaches either the upper or lower limit of the workspace.

Using this social learning technique, the robot was able to very quickly obtain a reasonably accurate behavior with very little training. While these results are somewhat qualitative, they do provide a behavior that is sufficiently believing to point to objects of reasonable size that are presented to the robot.

5.3.2 Trajectory Following for Mimicry

Pointing gestures move the end effector from a rest position to some point on the manifold defined by the other postural primitives. It is also possible to create arm trajectories that move along curves within that manifold. Given a set of positions

[2] Rather than having a human intervene, it would also be possible for the robot to simply use the visual movement of its arm as the target. In practice, discriminating the robot's hand from its forearm or elbow is a difficult visual task. Furthermore, this would imply a level of self-understanding that we have been unwilling to assume for this work.

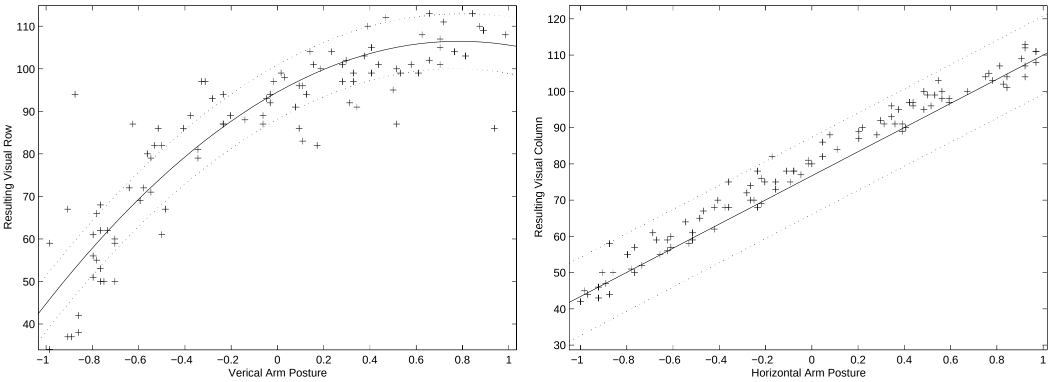


Figure 5-6: Resulting function fit for the pointing behavior using social training. Cross marks indicate training points obtained when the robot generated a random reach using postural primitive parameters shown on the x-axes and the resulting image coordinates of the most salient stimulus. Polynomial fitting by least-squares (solid line) and confidence intervals of 95% (dotted lines) are also shown.

within that manifold, a trajectory can be formed either by simply updating the current command posture at a rapid rate (30-100 Hz), or by interpolating a set of way points between positions for movements over longer time intervals.

Mapping from a set of points in the image plane to a set of arm postures can be done in a few different ways. One simple option is to map the image plane into the range [-1,1] along both dimensions by dividing each pixel location by one-half the height and width respectively. This has the effect of allowing the robot to match the range of its field of view to the range of its arm movements. This mapping will suffer from local distortions, but will preserve the general shape and direction of the trajectory of positions in the image plane. A second option is to recognize the extent of the visual trajectory and use that boundary as the full range of movement of the arm. For example, if the movement were coming from a person, it might be desirable to map the person's range of motion to the robot's range of motion. A third option is to use the mappings developed when learning to point to a visual target, which convert visual coordinates to arm postures, to map the visual trajectories to a trajectory of arm postures in the same localized region of space. This option is useful for attempting to match the robot's arm trajectories to objects in the real world. By following the mapping obtained when learning to point, the robot in effect points to and follows the visual impression of the object. In practice, we have used all three of these options. Chapter 9 describes the details on the effects of these different mappings on the behavior of the robot.

5.4 Conclusion

This chapter has presented a set of basic robot behaviors that have a close similarity to human behavioral counterparts. These behaviors give a basis upon which the perceptual and cognitive abilities discussed in the following chapters are based. There are certainly many additional improvements that could be made to these individual behavioral components to strengthen this foundation. Improvements to these basic behavioral systems would enrich the behavioral repertoire of the robot, but these basic behaviors are sufficient to demonstrate the effects of the embodied theory of mind model.

Chapter 6

Visual Attention

> *Seeing the world around you is like drinking from a firehose. The flood of information that enters the eyes could easily overwhelm the capacity of the visual system. To solve this problem, a mechanism – attention – allows selective processing of the information relevant to current goals.* – Kanwisher & Downing (1998, p. 57)

6.1 Introduction

A common problem for both animal and mechanical perceptual systems is that there are too few computational and motor resources to completely process all of the incoming perceptual signals. Attention is a mechanism for allocating these limited resources to the most relevant sensory stimuli. The most common view of attention in human psychophysics is that there are two levels of perceptual processing, pre-attentive and post-attentive, and that attention serves as a gateway to limit the amount of information that enters into the limited capacity of the post-attentive processes (Treisman, 1985). The pre-attentive systems are relatively simple computations that occur in parallel across the entire sensory stimulus (for example, across the entire retina for visual processing or across the entire tonal spectrum for auditory signals). Pre-attentive processing is automatic and not available to conscious inspection, although it can be influenced in limited ways by higher-level conceptual processes. Post-attentive processes make use of limited resources in memory, computational power, or motor responses. Due to these limited resources, post-attentive processes can only be applied in serial. Post-attentive processes are deliberate actions and are available to conscious inspection and planning. Attention mechanisms integrate influences from the pre-attentive mechanisms in order to direct post-attentive processes. These mechanisms are solving a saliency problem, that is, they are determining which stimuli out of the entire sensory scene are interesting and worthy of further attention.

This chapter discusses the construction of an attention system which directs limited computational resources and selects among potential behaviors by combining

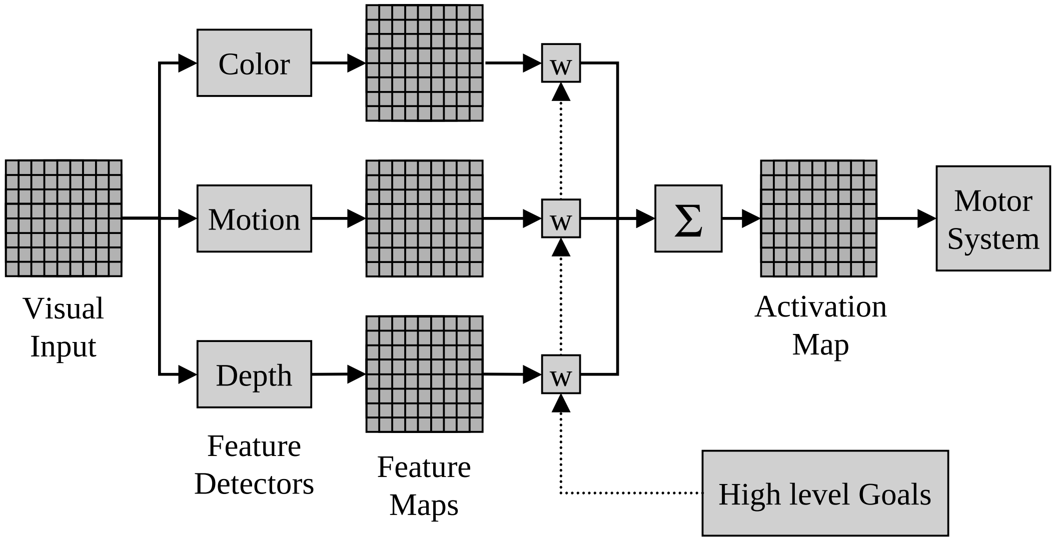


Figure 6-1: A model of human visual search and attention by Wolfe (1994). Visual stimuli are processed by a variety of low-level perceptual filters which produce individual feature maps, which are weighted and summed into a single activation map. Peaks in the activation map are areas of interest and are allocated limited computational or motor resources. High-level cognitive systems can influence the selection process only by modifying these weights.

perceptions from a variety of modalities with the existing motivational and behavioral state of the robot.[1] This is a critical ability both for maintaining behavioral consistency (which allows the human to have a more natural interaction) and for allowing the robot to operate in cluttered and noisy environments. The implementation is based upon models of human attention and visual search and has been a useful tool in predicting faults with existing models of visual attention. The implementation is opportunistic, deliberative, and socially grounded. The robot responds opportunistically to stimuli that an infant would find salient while also being able to perform deliberate search sequences based on high-level task constraints. Finally, the system is socially mediated in being able to respond to natural social cues that humans readily use to draw attention to a stimulus.

6.2 Implementation Overview

The implementation discussed here is based upon Wolfe's "Guided Search 2.0" model of human visual attention and visual search (Wolfe, 1994). This model integrates

[1] The original implementation of this work was carried out on Kismet in collaboration with Cynthia Breazeal (Breazeal & Scassellati, 1999). Since then, the architecture has been extended to include influences from joint reference behaviors (see chapter 10) and standardized for use on all three of our robot platforms.

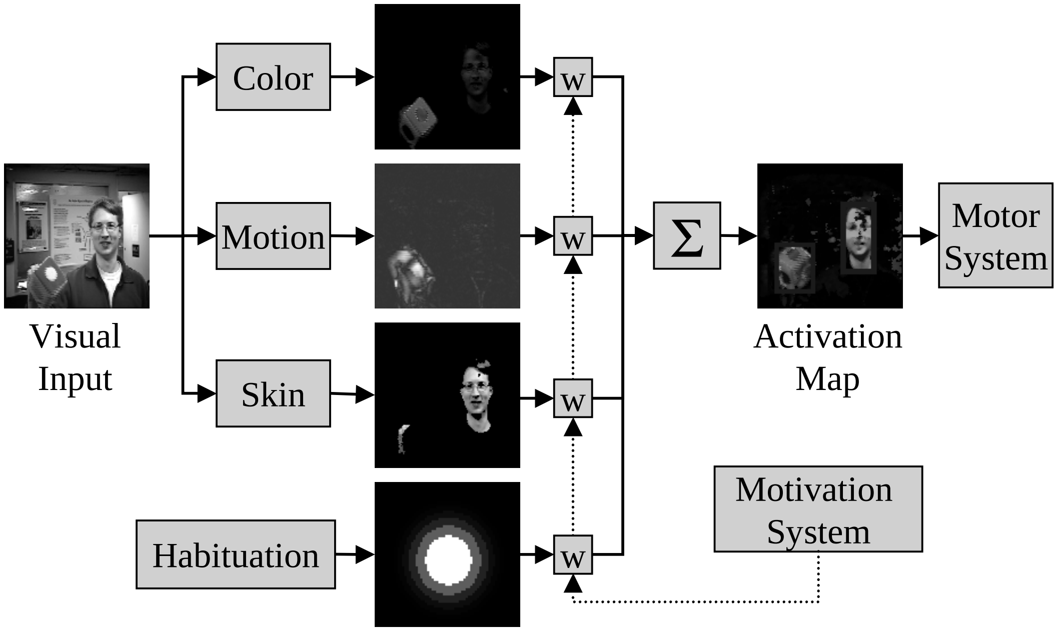


Figure 6-2: Overview of the attention system. A variety of visual feature detectors (color, motion, and face detectors) combine with a habituation function to produce an attention activation map. The attention process influences eye control and the robot's internal motivational and behavioral state, which in turn influence the weighted combination of the feature maps. Displayed images were captured during a behavioral trial session.

evidence from Treisman (1985), Julesz & Krose (1988), and others to construct a flexible model of human visual search behavior. In Wolfe's model (see figure 6-1), visual stimuli are filtered by broadly-tuned "categorical" channels (such as color and orientation) to produce *feature maps* in which high values indicate areas of interest. These feature maps may contain multiple categories of filtering. For example, the feature map for color may contain independent representations for red, yellow, green, and blue, each of which may contribute to the single color feature. These feature maps are retinotopically organized, maintaining the same 2-D projection as the retina. Individual feature maps are weighted and combined by point-wise summation to produce a single *activation map.* The peaks in this activation map indicate the most salient objects in the scene, and the scan path of an individual is computed by following the sequence of most activated regions. Top-down activation can drive a visual search by influencing the activation map through the weightings that are applied before summation. High-level processes may not have arbitrary effects on the visual search process; only through modifications of these weights may these processes find influence. For example, the search may be modified to be preferential towards "blue" stimuli or for "vertical" stimuli, but it may not execute arbitrary searches.

This model does well at explaining many of the conjunctive search effects noticed by Treisman (1985), Julesz & Bergen (1983), and Nakayama & Silverman (1986). For example, when presented with a field of objects consisting of a single circle among many squares, subjects are able to report the presence of the circle immediately. A similar effect is observed when searching for a blue item among red, or a vertical line among horizontal lines. These searches are simple and can be conducted in a constant amount of time that does not depend upon the number of distractors. To the observer, the query item seems to "pop-out" instantly from among the distractors. However, many conjunctive searches cannot be done in this fashion. For example, searching for a square red object among distractors that contain both circles and squares of both red and blue color takes an amount of time that increases linearly with the number of distractors. These complex searches do not result in "pop-out" effects and require an active search to find the query object. Wolfe's model explains these effects through the modifications that top-down processes may make on the summation weights. When looking for items that match particular categorical channels (such as "blue"), the weighting can influence the activation map to bring attention immediately to the query object. However, the model is incapable of performing more complex searches, which cannot be framed in terms of a single set of weight values. Thus, a search for objects that are both red and circular cannot be expressed by a single set of weights without also biasing the search toward objects that are merely red or merely circular. In this way, Wolfe's model provides a good estimation of the visual search behavior observed in humans.

This implementation does not attempt to match human performance exactly (a task that is difficult with current component technology), but rather requires only that the robotic system perform enough like a human that it is capable of maintaining a normal social interaction. Our implementation is similar to other models based in part on Wolfe's work (Itti et al., 1998; Hashimoto, 1998; Driscoll et al., 1998), but additionally operates in conjunction with motivational and behavioral models, with moving cameras, and it differs in dealing with habituation issues. The following sections will describe the low-level pre-attentive features that have been implemented, the habituation mechanisms that have been added, the methods for combining feature maps, and the ways in which high-level tasks influence this system.

6.3 Low-level Perceptual Systems

One reason that objects can become salient is because they have inherent properties that are intrinsically interesting. For example, objects that are brightly colored or moving naturally attract attention. The low-level perceptual systems process visual input directly to represent the inherent saliency of an object. The implementation described here focuses on three pre-attentive processes: color, motion, and skin color pop-outs.[2] Both color and motion are inherent properties that are recognized by Wolfe as part of his model. We have additionally added skin color as a pre-attentive filter to bias the robot toward attending to people. While this is a less well-supported

[2] In previous versions of this work, a face detector pop-out was used instead of the skin color filter. However, this complex processing was a less faithful representation of Wolfe's model, and also was more computationally expensive than the pre-attentive features should be.

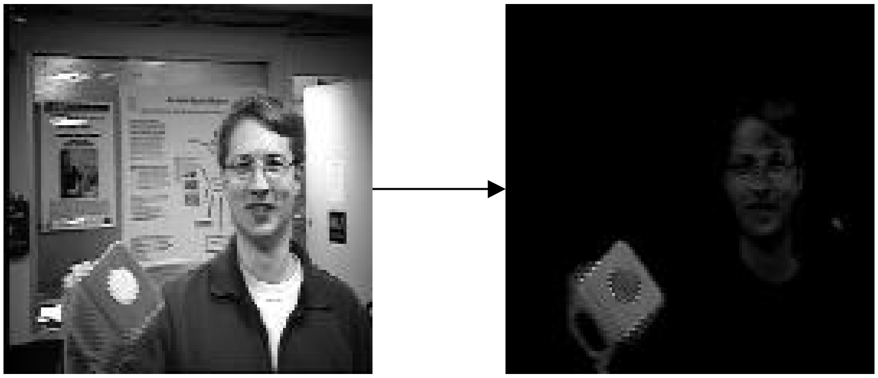


Figure 6-3: The color saliency feature detector. At left, the raw 128×128 image. At right, the feature map produced by processing the raw image with the color saliency feature detector.

assumption, there are reasons to believe that skin colors attract the attention of infants (Aslin, 1987).

Responses to these inherent properties represent opportunistic behaviors, that is, they are responses that are driven by the environment directly. The responses are also socially mediated, in that they respond to stimuli that are often used socially to indicate objects of interest. For example, when attempting to show an object to a child (or a robot), a person will often shake the object or present it by moving it closer. These simple social cues are recognized through these low-level filters and thus can influence all later stages of behavior.

6.3.1 Color Saliency Feature Map

One of the most basic and widely recognized visual features is color (see figure 6-3). Our models of color saliency are drawn from the complementary work on visual search and attention of Itti et al. (1998). The incoming video stream contains three 8-bit color channels (r, g, and b) which are transformed into four color-opponency channels (r', g', b', and y') to better match the human color-opponent center-surround photoreceptors. Each input color channel is first normalized by the luminance l (a weighted average of the three input color channels):

$$r_n = \frac{255}{3} \cdot \frac{r}{l} \qquad g_n = \frac{255}{3} \cdot \frac{g}{l} \qquad b_n = \frac{255}{3} \cdot \frac{b}{l}$$

These normalized color channels are then used to produce four opponent-color channels:

$$\begin{aligned} r' &= r_n - (g_n + b_n)/2 \\ g' &= g_n - (r_n + b_n)/2 \\ b' &= b_n - (r_n + g_n)/2 \end{aligned}$$

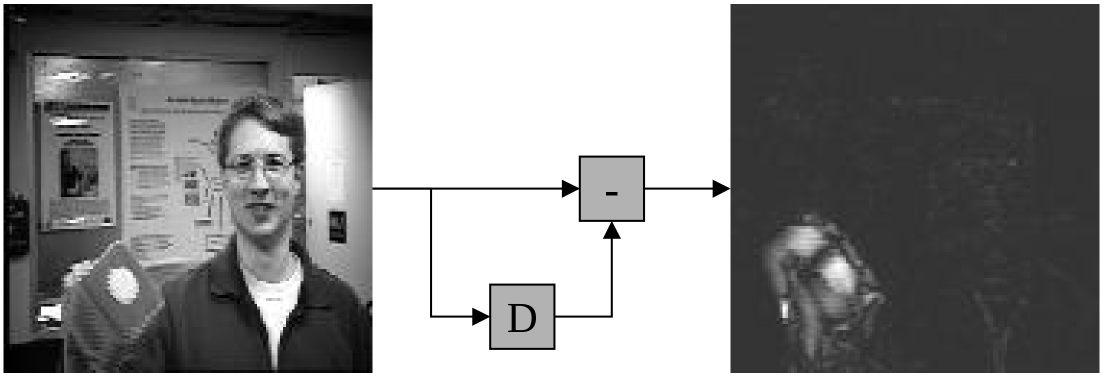


Figure 6-4: The visual motion feature detector. At left, the raw 128 × 128 image. At right, the feature map produced by subtracting the previous frame in the image sequence.

$$y' = \frac{r_n + g_n}{2} - b_n - \|r_n - g_n\|$$

The four opponent-color channels are clipped to 8-bit values by thresholding. While some research seems to indicate that each color channel should be considered individually (Nothdurft, 1993), we choose to maintain all of the color information in a single feature map to simplify the processing requirements (as does Wolfe (1994) for more theoretical reasons). The maximum of the four opponent-color values is computed and then smoothed with a uniform 5 × 5 field to produce the output color saliency feature map. This smoothing serves both to eliminate pixel-level noise and to provide a neighborhood of influence to the output map, as proposed by Wolfe (1994). A single computational node computes this filter and forwards the resulting feature map both to the attention process and a VGA display processor at a rate of 30 Hz. The processor produces a pseudo-color image by scaling the luminance of the original image by the output saliency while retaining the same relative chrominance (as shown in Figure 6-2).

6.3.2 Motion Saliency Feature Map

In parallel with the color saliency computations, a second processor receives input images from the frame grabber and computes temporal differences to detect motion (see figure 6-4). The incoming image is converted to grayscale and placed into a ring of frame buffers. A raw motion map is computed by passing the absolute difference between consecutive images through a threshold function $\mathcal{T}$:

$$M_{raw} = \mathcal{T}(\|I_t - I_{t-1}\|)$$

This raw motion map is then smoothed with a uniform 5 × 5 field. Additionally, the output of the motion filter is suppressed during certain eye movements. For 200

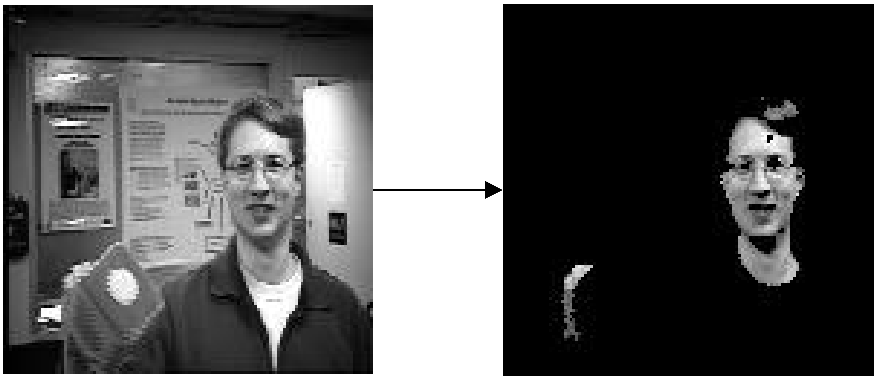


Figure 6-5: The skin color feature detector. At left, the raw 128 × 128 image. At right, the result of applying the skin color filter.

milliseconds following the onset of a saccade, the output of the motion detector is completely suppressed to allow for the image to stabilize. When performing smooth-pursuit tracking, the non-central areas of the peripheral image are suppressed from the output of the motion map. In general, this removes the motion blur of background objects from attracting the robot's attention (although those objects still maintain other inherent saliency properties). The motion saliency feature map is computed at 30 Hz by a single processor node and forwarded both to the attention process and the VGA display.

6.3.3 Skin Color Feature Map

The third pre-attentive feature detector identifies regions that have color values that are within the range of skin tones (see figure 6-5) (Breazeal et al., 2000*b*). Incoming images are first filtered by a mask that identifies candidate areas as those that satisfy the following criteria on the red, green, and blue pixel components:

$$2g > r > 1.1g \qquad 2b > r > 0.9b \qquad 250 > r > 20$$

These constants were determined by examining the clusters of skin pixels in hand-labeled images. The final weighting of each region is determined by a learned classification function that was trained on hand-classified image regions. The output is median filtered with a small support area to minimize noise. This detector also operates at 30 Hz.

6.4 Habituation

Wolfe's model explains reactions to static scenes, but does not model how the human responds dynamically over time. One simple mechanism that gives a realistic human-like response is to habituate to stimuli that are currently under attention. For our

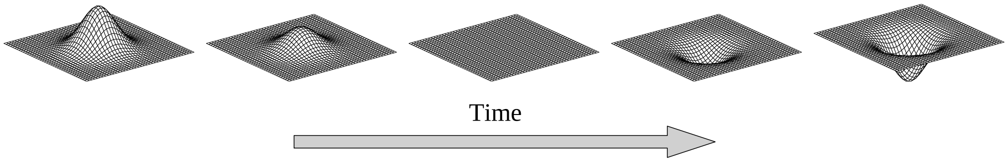


Figure 6-6: The habituation function is a Gaussian field with amplitude that decays by a linear time constant. The time decay is reset whenever the eyes move to acquire a new target of attention.

robot, the current object under consideration is always the object that is in the center of the peripheral visual field. This is extremely relevant on Cog and Lazlo, since the center of the peripheral field of view is also the area inside the foveal field of view. The habituation mechanism serves both to initially enhance an object when it first comes under attention and later to make the object less and less interesting.

The habituation function can be viewed as a feature map that initially maintains eye fixation by increasing the saliency of the center of the field of view and slowly decays the saliency values of central objects until a salient off-center object causes the eyes to move. The habituation function is a Gaussian field $G(x, y)$ centered in the field of view with $\theta = 30$ pixels (see figure 6-6). It is combined linearly with the other feature maps using the weight

$$w = W \cdot max(-1, 1 - \Delta t/\tau)$$

where w is the weight, Δt is the time since the last habituation reset, τ is a time constant, and W is the maximum habituation gain. Whenever the eyes move, the habituation function is reset, forcing w to W and amplifying the saliency of central objects until a time τ when $w = 0$ and there is no influence from the habituation map. As time progresses, w decays to a minimum value of $-W$ which suppresses the saliency of central objects. In the current implementation, we use a value of $W = 255$ (to remain consistent with the other 8-bit values) and and a time constant $\tau = 5$ seconds. The habituation function is treated as another low-level perceptual feature, and may have a weighting associated with it. This weighting allows for an amplification of the effects of the habituation signal with respect to the other feature maps.

6.5 Linking Feature Maps to Behaviors

The results from the individual feature maps are independent measurements of inherent object saliency. The attention system weights each of these features and combines them to produce a command signal for the eye movements.

6.5.1 Combining Feature Maps

Each of the feature maps contains an 8-bit value for each pixel location which represents the relative presence of that visual scene feature at that pixel. The attention process combines each of these feature maps using a weighted sum to produce an attention activation map. The gains for each feature map default to values of 200 for color, 40 for motion, 50 for skin color, and 10 for habituation. The attention activation map is thresholded to remove noise values, and normalized by the sum of the gains.

Pixels passing threshold are each assigned a tag which indicates the object area to which they belong. Whenever two pixels with different tags are adjacent (using 8-connectivity), those two tags are merged into a single tag. Once no further merges are possible, the bounding box and centroid of each tagged region are computed. To compensate for some of the noise properties of the camera system, if any two tagged regions have bounding boxes that overlap or that are within ϵ pixels of overlap, those two regions are also merged. While this second merge procedure limits the resolution of the labeling procedure, in practice this step was necessary to ensure the robustness of the system; many typical objects had two or more salient areas that were very close but were not consolidated by the 8-connectivity merge because of a few pixels of noise. In this implementation, a value of $\epsilon = 2$ was used.

Statistics on each region are collected, including the centroid, bounding box, area, average attention activation score, and average score for each of the feature maps in that region. The tagged regions that have an area in excess of 30 pixels are sorted based upon their average attention activation score. The attention process provides the top three regions to both the eye motor control system and the behavior and motivational systems.

The entire attention process (with habituation) operates at 30 Hz on a single processor node. The speed varies slightly (by less than 10%) with the number of attention activation pixels that pass threshold for region growing. While the implementation could be further optimized, these small deviations have little impact on the behavior of the system.

6.5.2 Attention Drives Eye Movement

The eye motor control process acts on the data from the attention process to center the eyes on an object within the visual field using the learned saccade behavior described in section 5.1.1. The centroid of the most salient region is used to determine the target of the saccade. Additionally, the attention activation score and the individual feature map scores of the most salient region are made available to higher level processes so that they may base behavioral choices on the target of interest.

Each time the eyes move, the eye motor process sends two signals. The first signal inhibits the motion detection system for approximately 200 milliseconds, which prevents self-motion from appearing in the motion feature map. The second signal resets the habituation state.

Because the effects of habituation are combined with the other low-level features,

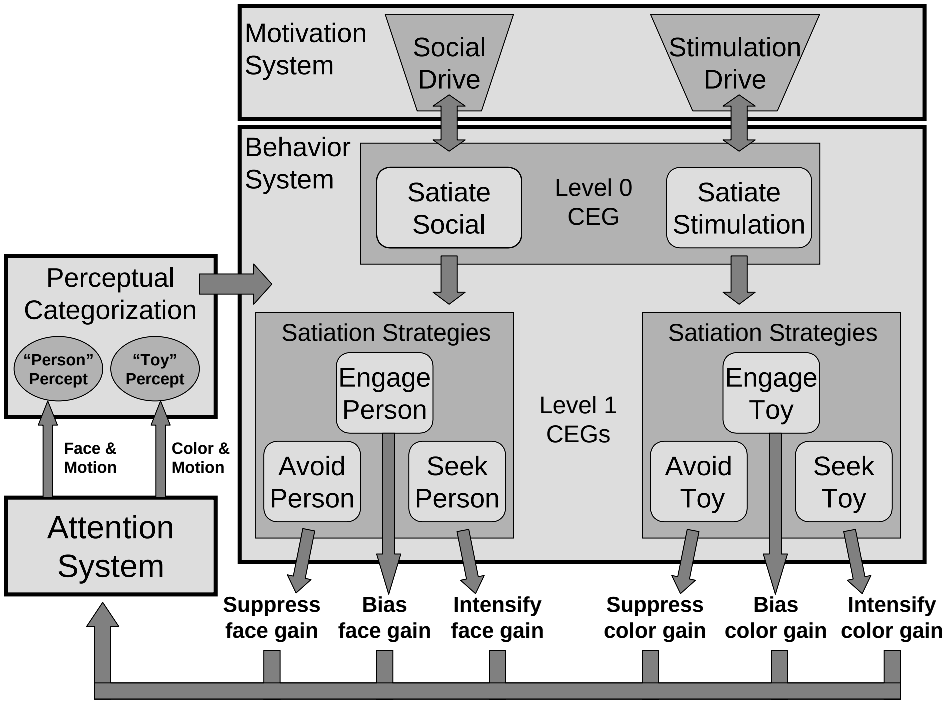


Figure 6-7: Influences of Kismet's motivation and behavior systems on the attention system. Kismet has two homeostatic drives; one drive is to interact socially with people and the other is to be stimulated with brightly colored toys. The drives influence the selection of behaviors at multiple levels of a cross-exclusion groups (CEG), allowing only one behavior at each level to activate its children. Each of the behaviors at the leaves can influence the weight gains of the attention system.

there is no fixed time period for which the robot will maintain attention on a target. The length of fixation will depend on the inherent saliency properties of the object, the decay rate of the habituation Gaussian, and the relative weights of the feature maps. In this way, the robot's behavior becomes richly enmeshed with the complexities of the environment.

6.6 Influences from High-Level Tasks

We have experimented with the mechanisms for allowing high-level processes to influence and modify the attention model using the robot Kismet. Kismet's behavior and motivation system have been described extensively (Breazeal & Scassellati, 2001*a*; Breazeal, 2000), but to understand the impact on the attention system, a brief sketch of the behavior and motivation system will be presented here.

The design of Kismet's motivation and behavior systems (modeled after theories of Lorenz (1973)) enable it to socially interact with a human while regulating the intensity of the interaction via expressive displays. Post-attentive perceptual processes classify stimuli into *social* stimuli (i.e., people, which move and have faces) which

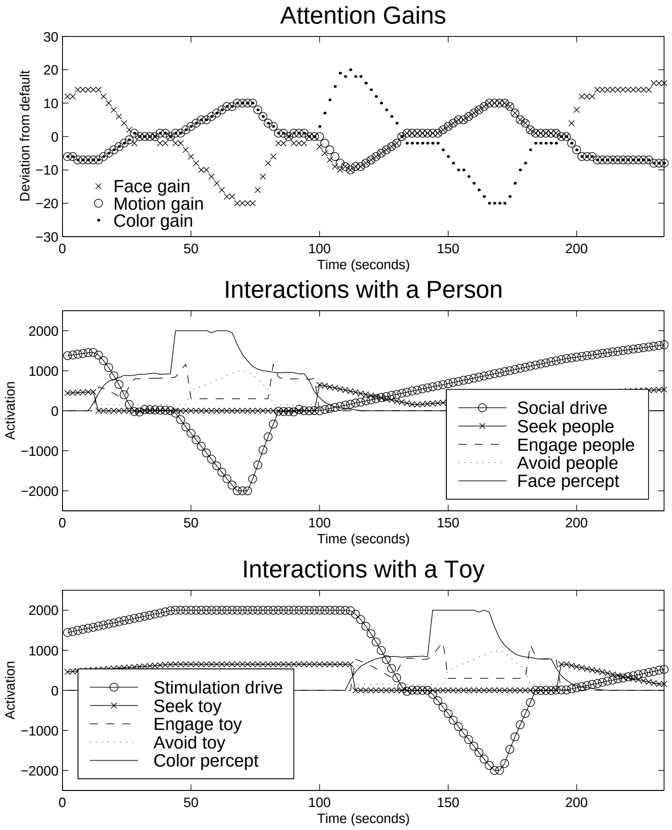


Figure 6-8: Changes of the motion, skin(face), and color gains from top-down motivational and behavioral influences (top). When the *social* drive is activated by face stimuli (middle), the face gain is influenced by the *seek people* and *avoid people* behaviors. When the *stimulation* drive is activated by colorful stimuli (bottom), the color gain is influenced by the *seek toys* and *avoid toys* behaviors. All plots show the same 4 minute period.

satisfy a drive to be social and *non-social* stimuli (i.e., toys, which move and are colorful) which satisfy a drive to be stimulated by other things in the environment. For each drive, there is a desired operation point, and an acceptable bounds of operation around that point (the *homeostatic regime*). As long as a drive is within the homeostatic regime, that corresponding need is being adequately met. Unattended, drives drift toward an under-stimulated regime. Excessive stimulation (too many stimuli or stimuli moving too quickly) push a drive toward an over-stimulated regime. Kismet's drives influence behavior selection by preferentially passing activation to select behaviors. By doing so, the robot is more likely to activate behaviors that serve to restore its drives to their homeostatic regimes.

As shown in Figure 6-7, the face gain is enhanced when the *seek people* behavior is active and is suppressed when the *avoid people* behavior is active. Similarly, the color gain is enhanced when the *seek toys* behavior is active, and suppressed when the *avoid toys* behavior is active. Whenever the *engage people* or *engage toys* behaviors are active, the face and color gains are restored to their default values, respectively.

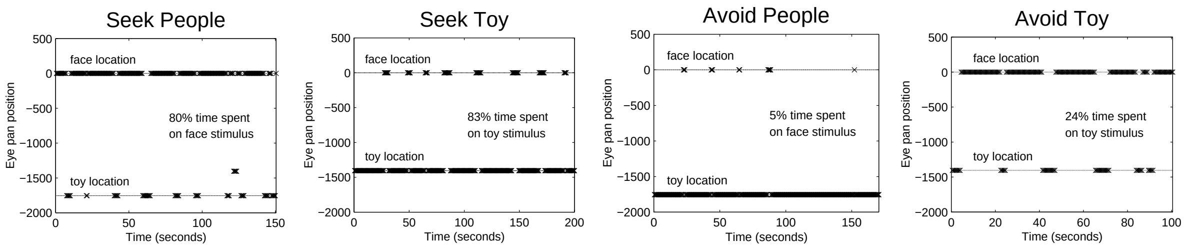


Figure 6-9: Preferential looking based on habituation and top-down influences. When presented with two salient stimuli (a face and a brightly colored toy), the robot prefers to look at the stimulus that has behavioral relevance. Habituation causes the robot to also spend time looking at the non-preferred stimulus.

Weight adjustments are constrained such that the total sum of the weights remains constant at all times. Figure 6-8 illustrates how the face, motion, and color gains are adjusted as a function of drive intensity, the active behavior, and the nature and quality of the perceptual stimulus.

6.6.1 Evaluating the Effects of Top-Down Influences

Top-down gain adjustments combine with bottom-up habituation effects to bias the robot's gaze preference (see Figure 6-9). When the *seek people* behavior is active, the face gain is enhanced and the robot prefers to look at a face over a colorful toy. The robot eventually habituates to the face stimulus and switches gaze briefly to the toy stimulus. Once the robot has moved its gaze away from the face stimulus, the habituation is reset and the robot rapidly re-acquires the face. In one set of behavioral trials when *seek people* was active, the robot spent 80% of the time looking at the face. A similar affect can be seen when the *seek toy* behavior is active — the robot prefers to look at a toy over a face 83% of the time.

The opposite effect is apparent when the *avoid people* behavior is active. In this case, the face gain is suppressed so that faces become less salient and are more rapidly affected by habituation. Because the toy is relatively more salient than the face, it takes longer for the robot to habituate. Overall, the robot looks at faces only 5% of the time when in this behavioral context. A similar scenario holds when the robot's *avoid toy* behavior is active — the robot looks at toys only 24% of the time.

Notice that in each of these cases, the influence of the high-level motivations are easily seen in the behavior of the system, but do not completely determine the behavior of the system. In this way, the robot is both deliberative in behaving according to these high-level goals and opportunistic in continuing to respond to stimuli that become salient through their inherent properties. The behavior of the system is based off both of these influences, and there is no fixed determination of the relative importance of low-level and high-level influences.

6.7 Summary

We have implemented a visual attention system based on models of human visual search and attention. The attention system is critically important for limiting the amount of information that must be processed, which allows the robot to operate in unstructured environments. This system combines information on inherent object properties (such as color saturation, motion, and skin color), high-level influences from motivations and goals, and a model of habituation to select objects in the visual scene that are socially relevant. In the next chapter, we begin to analyze the movement patterns of these objects in order to perform the most basic theory of mind task: the discrimination of animate from inanimate stimuli.

Chapter 7

The Theory of Body Module

> *Of course, algorithms for animateness and intentionality can lead to mistakes. They surely did not evolve in response to selection pressures involving two-dimensional figures moving across computer screens. These inhabitants of flatland just happen to fall within the actual domains to which the modules for animacy and intentionality spontaneously extend, as opposed to the proper domains for which the modules evolved (i.e., animate beings and intentional agents).* – Atran (1998, p. 555)

One of the most basic visual tasks for any organism is to distinguish between animate objects, which might be potential predators or mates, and inanimate objects. The distinctions between "alive" and "not-alive" are complex conceptual constructs that change drastically as children acquire new knowledge and reasoning capabilities (Keil, 1995; Carey, 1995; Gelman et al., 1983). While the discrimination of animate from inanimate certainly relies upon many distinct properties, including the object's texture, color, shape regularity, and perhaps symmetry, as well as the context of the observed object, Michotte (1962) and a host of others (for a review, see Scholl & Tremoulet, 2000) have demonstrated that animacy can be elicited by the movement of single points of light or simple geometric objects moving across a blank background. As Leslie (1982) and Cohen & Amsel (1998) observed, these basic spatial and temporal properties are recognized by children as early as six months of age.

In Leslie's model (1984), this discrimination is performed by the theory of body module (ToBY). ToBY uses a set of naive rules about how inanimate objects move through the world in order to classify inanimate from animate stimuli based purely on the spatial and temporal qualities of the object's movement. The rules that ToBY encapsulates are a somewhat simplified view of Newtonian physics in an environment with high levels of friction. These rules do not match the real mechanics of object motion, rather they represent our naive understanding of how objects move. Chaput & Cohen (2001) have begun to outline a connectionist architecture for interpreting these events as causal indicators. Their system uses the most primitive of perceptual data (single points) to develop causal explanations of collision events similar to those described by Michotte (1962) and shown in figure 3-3. Unlike the work presented

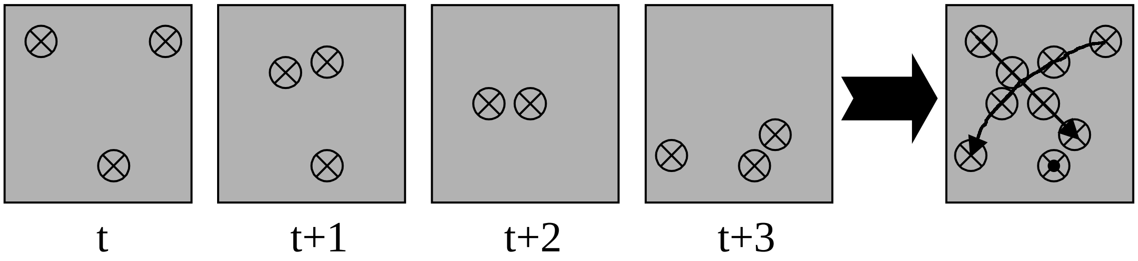


Figure 7-1: Outline of the motion correspondence problem. Each image frame in a video sequence contains some number of target locations which must be linked together to form spatio-temporal object trajectories.

here, their system deals with very clean perceptual signals and does not attempt to ground the perceptual data to real sensory systems. Their architecture also focuses almost exclusively on collision events and ignores many other interesting cognitive perceptual events. However, their work is complementary to this work in that it offers an example of how these naive physical laws might be learned autonomously.

Although most of the testing on the animate/inanimate distinction has been performed on simple geometric shapes on a flat screen, the discrimination that ToBY performs must operate on real-world stimuli. To provide the perceptual grounding for ToBY, salient objects generated by the attention system are linked together using a motion correspondence algorithm to form trajectories, which in turn serve as the inputs to ToBY (section 7.1). These trajectories are then processed by a multi-agent representation that mimics Leslie's ToBY module by attempting to describe trajectories in terms of naive physical laws (section 7.2). The results of the implemented system on real-world environments are introduced, and a comparison against human performance on describing identical data is discussed in section 7.3.

7.1 Computing Motion Trajectories

In order to classify object movement according to animacy, the ToBY module requires as input the trajectory of an object as it moves through space and time. However, the attention system has been designed to operate on single images. The output of the attention system is a set of object locations and properties for each image frame in the video sequence. The first problem that must be addressed is how to link these individual points together to form trajectories (see figure 7-1). This problem, often called trajectory formation or motion correspondence, has been extensively studied in the fields of target tracking and surveillance (Bar-Shalom & Formtann, 1988). Simple solutions track feature points between frames using a nearest-neighbor judgment (Tomasi & Kanade, 1992), or assume that the number of trajectories is a known constant (Chang & Aggarwal, 1991). However, these simple methods fail when dealing with trajectories that cross paths, when the number of trajectories changes dynamically, or when targets enter or leave the scene – all cases that regularly occur

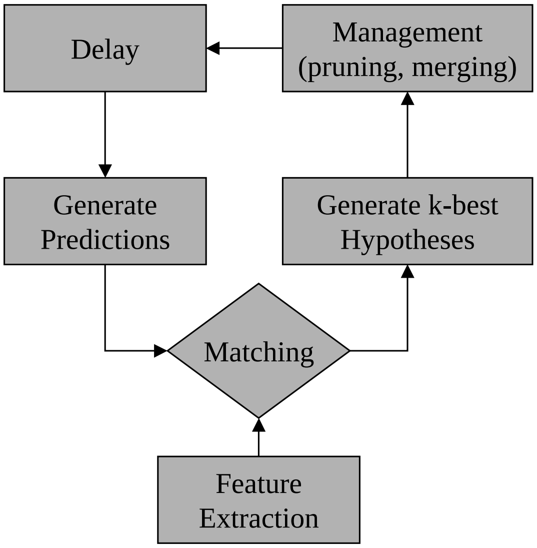


Figure 7-2: Flowchart for the multiple hypothesis tracking algorithm developed by Reid (1979) and implemented by Cox & Hingorani (1996).

in the visual environments that our robots inhabit. Furthermore, these methods tend to be sensitive to spurious measurements (noise in the target selection process) and often violate uniqueness constraints by assigning the same target position to multiple trajectories.

To address these problems, Reid (1979) proposed an algorithm called multiple hypothesis tracking, which was later implemented and evaluated by Cox & Hingorani (1996). At each timestep, the attention system produces a set of at most b salient objects. The centroids of those salient objects define a set of measurement points $\{P_t^1, P_t^2, ...P_t^b\}$ in each frame t. Given an incoming stream of these measurements, the objective of the multiple hypothesis tracking algorithm is to produce a labeled trajectory which consists of a set of points, at most one from each frame, which identify a single object in the world as it moves through the field of view:

$$T = \{P_1^{i_1}, P_2^{i_2}, ...P_t^{i_n}\}$$

The algorithm (see figure 7-2) operates by maintaining a set of hypotheses, each of which represents a possible trajectory for a subset of the total image points. As new measurements arrive, they are matched against existing hypothetical trajectories. These new measurements might extend an existing trajectory, start a new trajectory, or be classified as a false alarm (a sensory value that is noise, and should not be considered part of any particular trajectory). For each new data point, the algorithm generates all possible trajectory extensions, creations, and false alarms. The algorithm then does some hypothesis management by eliminating or merging duplicate trajectories, removing old trajectories, and pruning the set of trajectories by selecting the k best hypotheses and discarding the rest. The surviving hypotheses are used to generate predictions of where the next set of measurements should appear. The prediction of new feature points is based on a Kalman predictor, which uses the previous positions as internal state in order to predict velocity measurements and future positions. The predictions are then used as the basis upon which the matching of new features occurs.

The matching of features can be carried out in many ways. One possibility is simply to use the distance between the centroids of each target. In this option, individual feature points are matched to those that maintain a close spatial distance. This mechanism differs slightly from a generic nearest-neighbor matching, since an optimal hypothesis across multiple frames may have some non-optimal matches on particular frames. In other words, to get a better global solution a non-optimal local match may be accepted. More accurate and detailed matching metrics can be obtained by using some of the additional feature information associated with each of the targets generated by the attention system. For each target point produced by the attention system, the following information is available: centroid position, bounding box, total pixel area, total pixel saliency, total saliencies for each individual feature map (color, motion, skin, etc.), and a measurement of whether the object was within the current disparity plane. Matching can be done with any subset of these features by defining a statistical model of how the individual feature components are likely to change over time. In practice, the area, centroid position, and saliency components from the individual feature maps are used to evaluate the match criteria. Each feature is considered to have a normal distribution with variance determined empirically from a few hand-labeled trajectories.

The implementation of the multiple hypothesis tracking algorithm was based on code kindly provided by Ingemar Cox. Modifications of the existing code were required to transform the algorithm from a batch-processing implementation to a continuous on-line implementation. Additional modifications were implemented to embed the algorithm within the real-time process model and inter-process communication mechanisms used throughout the current implementation. The completed implementation runs in real-time (30 Hz) with a maximum of $b = 8$ measurement points in each frame and a maximum of $k = 300$ global hypotheses.

7.2 Naive Physics Experts

To implement the variety of naive physical laws encompassed by the Theory of Body module, a simple expert-based approach was chosen. Each expert represents knowledge of a single theory about the behavior of inanimate physical objects. For every trajectory T, each expert a computes both an animacy vote α_{Ta} and a certainty ρ_{Ta}. The animacy votes range from $+1$ (indicating animacy) to -1 (indicating inanimacy), and the certainties range from 0 to 1. For these initial tests, five experts were constructed: a static object expert, a straight line expert, an acceleration sign change expert, an elastic collision expert, and an energy expert. These experts were chosen to handle simple, common motion trajectories observed in natural environments and do not represent a complete set. Most notably absent are experts that recognize

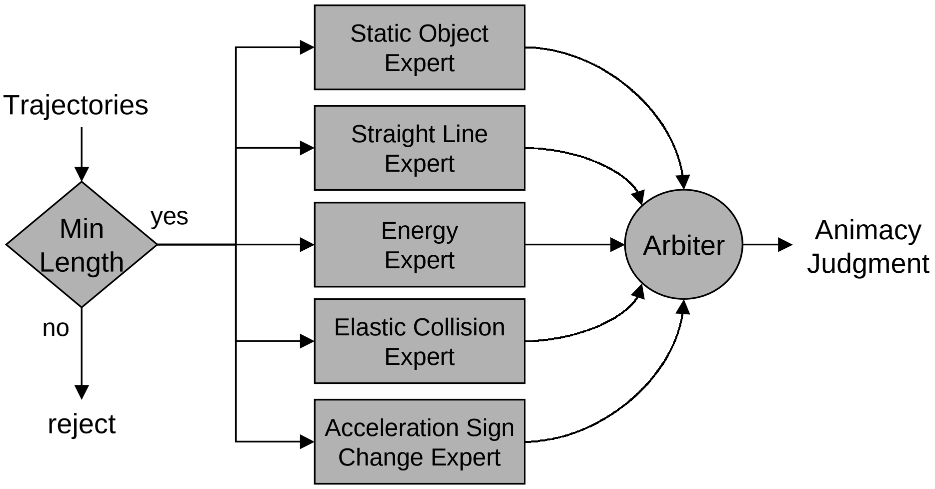


Figure 7-3: The architecture for the theory of body module.

repetitive motions as inanimate.

At each time step, every trajectory that passed a minimum length requirement was processed by each of the ToBY experts (see figure 7-3). The minimum length requirement was imposed to ensure that all trajectories contained sufficient data to compute statistical information against the noise background. Any trajectory with fewer than one-twentieth the maximum trajectory length or fewer than three data points is given an animacy vote $\alpha = 0.0$ with a certainty value of 1.0. In practice, maximum trajectory lengths of 60-120 were used (corresponding to trajectories spanning 2-4 seconds), so any trajectory of fewer than 3-6 data points was rejected. All trajectories that passed this test were evaluated by each ToBY expert, and the votes from each of these experts were tallied. Three different voting arbitration algorithms were tested to produce the final vote V_T for each trajectory T. The first voting method was a simple winner-take-all vote in which the winner was declared to be the expert with the greatest absolute value of the product:

$$V_T = \max_a \|\alpha_{Ta} \times \rho_{Ta}\|$$

The second method was an average of all of the individual vote products:

$$V_T = \frac{1}{A} \sum_a (\alpha_{Ta} \times \rho_{Ta})$$

where A is the number of experts voting. The third method was a weighted average of the products of the certainties and the animacy votes:

$$V_T = \frac{1}{A} \sum_a (w_a \times \alpha_{Ta} \times \rho_{Ta})$$

where w_a is the weight for expert a. Weights were empirically chosen to maximize performance under normal, multi-object conditions in natural environments and were kept constant throughout this experiment as 1.0 for all experts except the static object expert which had a weight of 2.0. The animacy vote at each time step is averaged with a time-decaying weight function to produce a sustained animacy measurement.

7.2.1 Static Object Expert

> *Objects that are stationary are inanimate.*

Because the attention system still generates target points for objects that are stationary, there must be an expert that can classify objects that are not moving as inanimate. The static object expert rejects any trajectory that has an accumulated translation below a threshold value as inanimate. The certainty of the measurement is inversely proportional to the translated distance and is proportional to the length of the trajectory.

7.2.2 Straight Line Expert

> *Objects that move in a straight line with constant deceleration are inanimate.*

The straight line expert looks for constant, sustained velocities. This expert computes the deviations of the velocity profile from the average velocity vector. If the sum of these deviations fall below a threshold, as would result from a straight linear movement, then the expert casts a vote for inanimacy. Below this threshold, the certainty is inversely proportional to the sum of the deviations. If the sum of the deviations is above a secondary threshold, indicating a trajectory with high curvature or multiple curvature changes, then the expert casts a vote for animacy. Above this threshold, the certainty is proportional to the sum of the deviations.

7.2.3 Elastic Collision Expert

> *Objects that rebound from a contact with another object in the same depth plane are inanimate.*

A collision in which the kinetic energy is the same before and after the collision is called elastic. In an inelastic collision, some of the kinetic energy is lost in friction or as heat as the shape or structure of the objects change due to the impact force. While most macroscopic interactions between objects are unlikely to be elastic, there are certain real-world events that look enough like elastic collisions that they represent an interesting domain. For example, a ball bouncing on the floor can be roughly modeled as an elastic collision. In terms of describing animacy, elastic collisions serve well as indicators of inanimacy, while inelastic collisions are not good estimators of either animacy or inanimacy (Premack, 1990). An inelastic collision could be the result of two inanimate objects striking each other and deforming (as would occur if a boulder rolling down a hill were to strike a tree), the result of an animate agent interacting with an inanimate object (such as a man stopping to rest against a tree), or the result of two animate agents interacting (such as two people stopping in the hall to chat).

One method for detecting elastic collisions would be to look at the interactions between all pairs of objects in the scene. This is not feasible for most real-world implementations, as the number of objects in the scene is potentially very large and often unknown. Furthermore, objects of interest may collide with objects that are often unrepresented (such as the floor). Rather than matching between pairs of salient objects, the elastic collision expert monitors each individual trajectory for potential collision points using a two-step method that does not explicitly represent the second object. For each salient trajectory (β), points of possible collisions are detected by watching for changes in the direction of the velocity vector by more than 90 degrees. Second, possible collision points are evaluated to determine if a solid object is located along the direction of the original velocity and within close spatial proximity to the point of contact. The elastic collision expert projects along the original velocity direction for a short distance ($\epsilon = 4$ pixels) and compares the result of the depth map at that location with the depth value of the trajectory. (Recall from section 5.1.3 that the depth map is computed using a simple correlation matching method on image patches). If an object is detected at the same (rough) depth, the collision is declared to be elastic. If all collisions for a trajectory are elastic, the trajectory is given an animacy vote of $\alpha = -1$ with a certainty of 0.50. The certainty value was chosen empirically to reflect the inaccuracies of the depth process.

7.2.4 Energy Expert

Objects that trade potential for kinetic energy are inanimate.

Bingham et al. (1995) have proposed that human adults judge animacy based on models of potential and kinetic energy. To explore their hypothesis, a simple energy model expert was implemented. The energy model expert judges an object that gains energy to be animate. The energy model computes the total energy of the system E based on a simple model of kinetic and potential energies:

$$E = \frac{1}{2}mv_y^2 + mgy$$

where m is the mass of the object, v_y the vertical velocity, g the gravity constant, and y the vertical position in the image. In practice, since the mass is a constant scale factor, it is not necessary for these calculations. This simple model assumes that an object higher in the image is further from the ground, and thus has more potential energy. The vertical distance and velocity are measured using the gravity vector from a three-axis inertial system as a guideline, allowing the robot to determine "up" even when its head is tilted. The certainty of the vote is proportional to the measured changes in energy.

7.2.5 Acceleration Sign Change Expert

Objects that often reverse their acceleration are animate.

One proposal for finding animacy is to look for changes in the sign of the acceleration. According to this proposal, anything that can alter the direction of its acceleration must be operating under its own power (excluding contact with other objects). The acceleration sign change expert looks for zero-crossings in the acceleration profile of a trajectory. Anything with more than one zero-crossing is given an animacy vote with a certainty proportional to the number of zero crossings.

7.3 Performance Evaluation

The performance of the individual experts was evaluated both on dynamic, real-world scenes at interactive rates and on more carefully controlled recorded video sequences.

For interactive video tasks, at each time step five attention targets were produced. Trajectories were allowed to grow to a length of sixty frames, but additional information on the long-term animacy scores for continuous trajectories were maintained as described in section 7.2. All three voting methods were tested. The winner-take-all and the weighted average voting methods produced extremely similar results, and eventually the winner-take-all strategy was employed for simplicity. The parameters of the ToBY module were tuned to match human judgments on long sequences of simple data structures (such as were produced by static objects or people moving back and forth throughout the room).

Figure 7-4 shows three sample trajectories taken during the unconstrained dynamic testing. Each of the samples shows a single image frame and the overlayed trajectories from the past 30 frames. In the first sequence (shown at left), five trajectories were detected. Four of the five were static objects (the author's head, a desk chair, a door in the background and a couch in the background) and were detected as inanimate by ToBY. The fifth was from the author's hand making a circular movement, which ToBY judged to be animate. In this example, the energy expert, acceleration sign change expert, and the straight line expert all voted that the hand trajectory was animate while the static object expert and the collision expert voted with zero certainty. This example was characteristic of many human-generated movements that were observed as people (both experienced and naive) interacted with the robot or simply passed by the robot. In the second sequence (shown at center), the author launched a wheeled office chair across the floor with a gentle push. The chair moved approximately 1.5 meters before it rolled to a stop. Of the five experts, two voted for inanimacy (the straight line expert and the acceleration sign change expert) and three gave an uncertain vote ($\alpha = 0$ for the remaining three experts). This trajectory was characteristic of many examples of manipulated objects in the environment

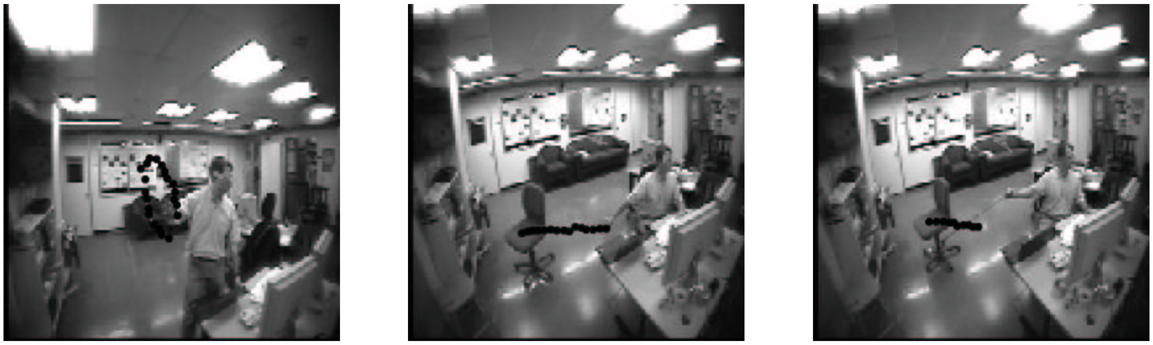


Figure 7-4: Three examples taken from the dynamic real-world tests of the ToBY module. Each example shows the overlayed trajectories as a connected line of black dots. At left, a hand making a circular movement was found to be animate. A desk chair that was launched across the floor (center) was inanimate, but the same desk chair that was moved across the floor by pushing it with an attached rod was animate (right). See text for further details.

including dropped books, desks and chairs pushed across the floor, and balls rolling on the floor. In the final example (at right), a long pole was attached to the office chair so that it could be pushed across the floor at varying speeds. The pole was too thin and neutrally-colored to be detected by the attention system as a target. In an attempt to reproduce inanimate motion, the author pushed the desk chair across the floor by pushing and pulling on the pole. In a few cases like this, the chair's trajectory was judged to be inanimate by the ToBY experts. However, in most cases, the ToBY experts judged the movement to be animate with two experts voting for animacy (the straight line expert and the acceleration sign change expert) and three experts voting with zero certainty. While it was difficult to discriminate between these two cases from the overlayed trajectory data, looking at the raw motion or at the individual components of the trajectory vectors showed clear differences between the natural chair movement and the pole-induced movement. In this case, the author was unable to mimic the natural movements of the chair rolling across the floor with sufficient fidelity to fool the ToBY classifier. While this evidence is anecdotal, it does demonstrate some of the power of this classification system. The following section provides more controlled comparisons of the ToBY experts against human judgments and the ground truth animacy of a variety of stimuli.

7.3.1 Motion Trajectory Stimuli

To further quantify the performance of the ToBY system and to evaluate the contributions of the individual experts, a set of dynamic motion data was recorded from the robot's attention system. While the robot was observing natural objects in both normal situations and planned sequences, the output of the attention system was written to a file. To accurately evaluate the performance of the ToBY experts, we desired to evaluate the machine results against both the ground truth of the visual scene and

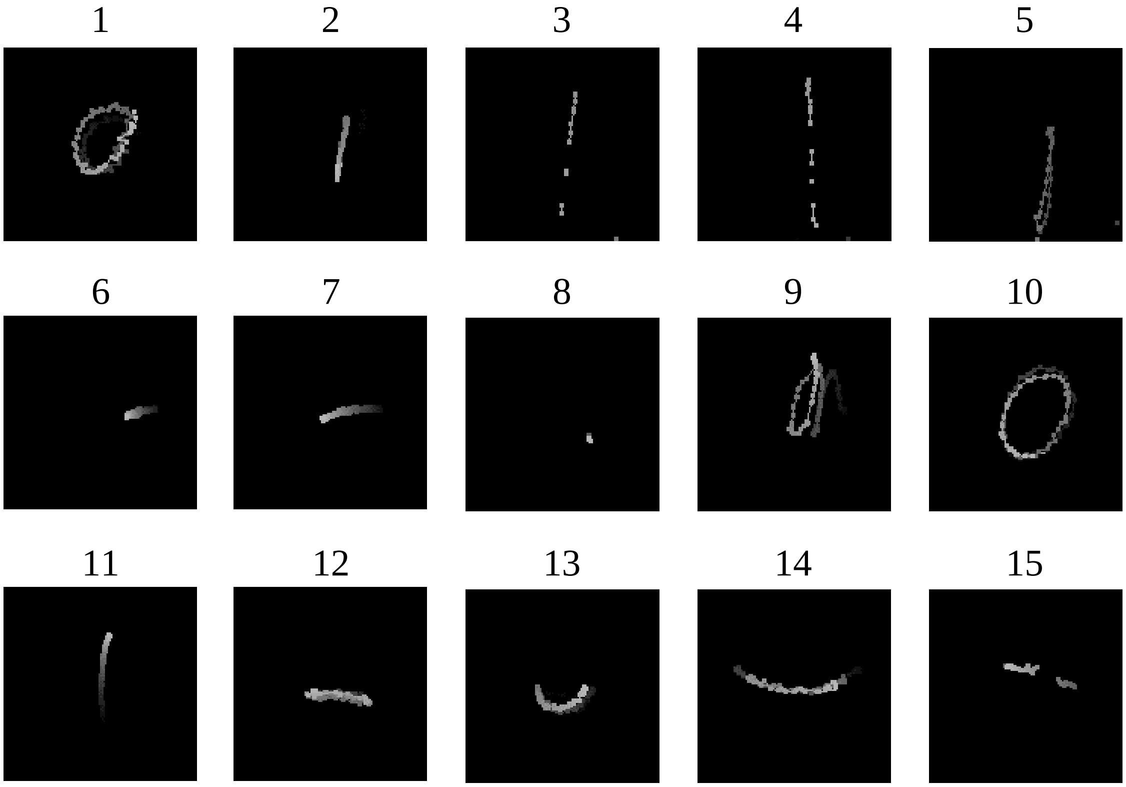


Figure 7-5: Fifteen stimuli used in the pilot study for evaluating ToBY. Each image is a static representation of a 2-4 second movie sequence in which a single point of light moved against a black background. The most recent position of the point of light is shown as a bright white spot, while older positions are less bright.

against what humans would judge given similar spatio-temporal information. To do this, a sequence of short movies were created from the attention data in which all potential contextual cues were removed. The human was shown only a single moving spot of light against a black background, in essence, the exact data that the ToBY system has available for an individual trajectory. The location of this dot was tied to the centroid location of the most salient object detected by the attention system. While the attention data contained multiple trajectories, only the trajectory of the most salient object was displayed to simplify in the collection of human animacy judgments and to simplify the reporting. Because each expert currently treats each trajectory independently, this restriction should not bias the comparison.

Two sets of data were collected using this technique. The first set of data contained fifteen sequences, each of which was two seconds long and contained only a single moving object. Each image in figure 7-5 shows a static representation of the moving stimulus in which the location of the moving dot is seen as a blur with the most recent positions drawn as the brightest points. This data was used to develop a reliable testing method for human subjects. Subjects were directed to a web page which contained instructions on how to perform the survey. Subjects were allowed to

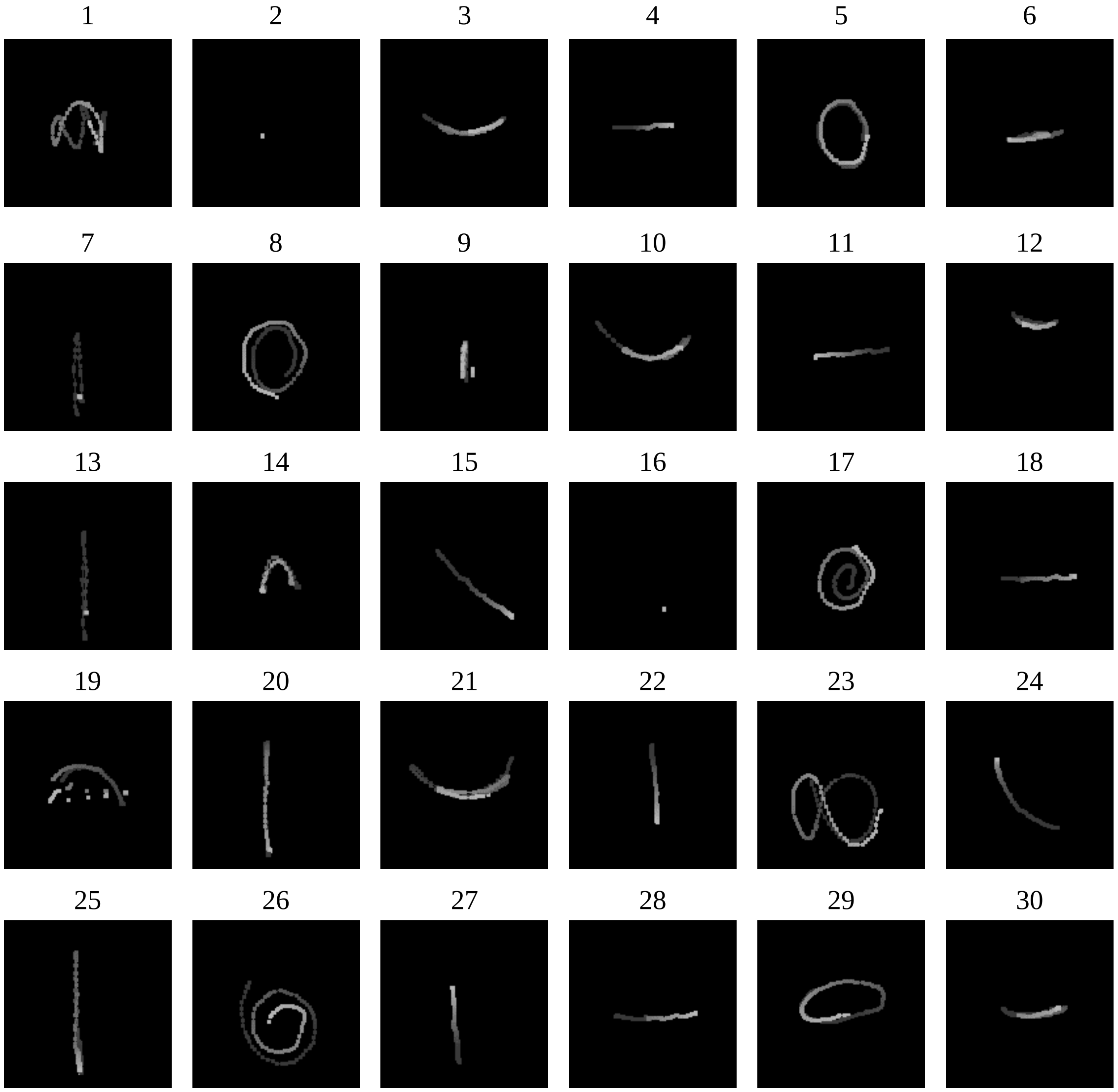


Figure 7-6: Thirty stimuli used in the evaluation of ToBY. Stimuli were collected by recording the position of the most salient object detected by the attention system when the robot observed natural scenes similar to the one shown in figure 7-4. Each image shown here is the collapsed sequence of video frames, with more recent points being brighter than older points. Human subjects saw only a single bright point in each frame of the video sequence.

view each 2-4 second movie sequence as many times as they wished, and afterward were asked to rate the animacy of the video sequence on a scale of 1 to 10, where 1 represented that they were certain that the object was animate and 10 represented certainty that the object was inanimate. Of the twenty subjects shown this initial pilot, five subjects found the initial set of instructions to be ambiguous, and one subject was confused on the recording method. Following the observations of these subjects and the critiques of three anonymous reviewers, the data from this pilot task was discarded and a slightly altered questionnaire was provided to a new set of subjects.

A second set of data consisting of thirty video segments of approximately 120 frames each was collected (see figure 7-6). A wide variety of movements were produced by putting real-world objects in front of the robot and recording the attention systems results. These trajectories included static objects (e.g., #2), swinging pendula (e.g., #3), objects that were thrown into the air (e.g., #7), as well as more complicated trajectories (e.g., #1). Figure 7-7 lists the trajectories grouped according to the category of movement and can be matched to figure 7-6 using the stimulus number in the second column. While many stimuli can easily be interpreted from the static representations shown in figure 7-6, a few deserve additional comment. Among the thrown objects, stimuli #7 and #13 are a result of normal objects being thrown into the air and falling, while stimuli #20 and #25 have been artificially created by an experimenter. In stimulus #20, the object is thrown into the air, caught near the apex of its ascent, held stationary momentarily, and then released. In stimulus #25, an experimenter moved a toy block through an upward motion followed by a downward motion but attempted to reverse the normal velocity profile by beginning with a slow movement, increasing the velocity as the block approached the apex, and slowing as the block descended.

7.3.2 Human Animacy Judgments

Thirty-two adult, volunteer subjects were recruited for the study using the second stimulus set. No subjects from the pilot study were allowed to participate in this experiment. Subjects ranged in age from 18 to 50, and included 14 women and 18 men. Subjects participated in a web-based questionnaire and were informed that they would be seeing video sequences containing only a single moving dot, and that this dot represented the movement of a real object. They were asked to rank each of the thirty trajectories shown in figure 7-6 on a scale of 1 (animate) to 10 (inanimate). Following initial pilot subjects (not included in this data), subjects were reminded that inanimate objects might move (such as a boulder rolling down a hill) but should still be treated as inanimate. Subjects received the following instructions on animacy judgments:

> To make these movies, a small light was attached to an object. The lights in the room were then turned off so that the only thing that could be seen was that single point of light. Your job will be to guess whether the object in each video was an animate object (such as a person's hand or

Stimulus Category	Stimulus Number	Notes
Static Objects	2	Stationary toy ball
	16	Stationary person
Thrown Objects	7	Ball is thrown into the air and falls
	13	Toy block is thrown into the air and falls
	20	Thrown object is held stationary at apex
	25	Velocity increases near apex
Circular Movements	5	Toy ball moving in a circle
	8	Person's hand moving in a circle
	17	Hand that spirals inward
	26	Slow inward spiral
	29	Elliptical movement
Straight Line Movements	4	Toy ball moves left to right
	11	Toy block moves right to left
	22	Hand moving slowly downward
	27	Hand moving slowly upward
	15	Ball rolling down an inclined plane
	24	Ball being pulled up an inclined plane
Pendula	3	Movement along a large radius
	10	Initial release has longer arc
	21	Object "settles" quickly
	30	Almost horizontal movement
	12	Rapid alternation of direction
Erratic Movements	1	Random movements created by a moving person
	6	Rapid left/right alternating movements
	9	Object bounces up and down
	14	Repeated left/right hops
	18	Left/right movement starts, stops, starts again
	19	Tracking system failure
	23	"Figure eight" movement
	28	Left/right movement starts, stops, starts again

Figure 7-7: Description of the stimuli from figure 7-6, grouped by motion class.

> a dog) or an inanimate object (such as a book falling off a shelf or a ball rolling across the floor). You will be asked to rate each object on a scale of 1 (animate) to 10 (inanimate).

Subjects were allowed to review each video sequence as often as they liked, and no time limit was used.

The task facing subjects was inherently under-constrained, and the animacy judgments showed high variance (a typical variance for a single stimulus across all subjects was 2.15). Subjects tended to find multiple interpretations for a single stimulus, and there was never a case when all subjects agreed on the animacy/inanimacy of a trajectory. To simplify the analysis, and to remove some of the inter-subject variability, each response was re-coded from the 1-10 scale to a single animate (1-5) or inanimate (6-10) judgment. Subjects made an average of approximately 8 decisions that disagreed with the ground truth values. This overall performance measurement of 73% correct implies that the task is difficult, but not impossible. Column 4 of figure 7-8 shows the percentage of subjects who considered each stimulus to be animate. In two cases (stimuli #13 and #9), the majority of human subjects disagreed with the ground truth values. Stimulus #9 showed a dot moving alternately up and down, repeating a cycle approximately every 300 milliseconds. Subjects reported seeing this movement as "too regular to be animate." Stimulus #13 may have been confusing to subjects in that it contained an inanimate trajectory (a ball being thrown and falling) that was obviously caused by an animate (but unseen) force.

7.3.3 ToBY Animacy Judgments

The identical video sequences shown to the human subjects were processed by the trajectory formation system and the ToBY system. Trajectory lengths were allowed to grow to 120 frames to take advantage of all of the information available in each short video clip. A winner-take-all selection method was imposed on the ToBY experts to simplify the reporting of the results, but subsequent processing with both other voting methods produced identical results. The final animacy judgment was determined to be the winning expert on the final time step. Columns 6 and 5 of figure 7-8 show the winning expert and that expert's animacy vote respectively.

Overall, ToBY agreed with the ground truth values on 23 of the 30 stimuli, and with the majority of human subjects on 21 of the 30 stimuli. On the static object categories, the circular movement stimuli, and the straight line movement stimuli, ToBY matched the ground truth values perfectly. This system also completely failed on all stimuli that had natural pendulum-like movements. While our original predictions indicated that the energy expert should be capable of dealing with this class of stimuli, human subjects seemed to be responding more to the repetitive nature of the stimulus rather than the transfer between kinetic and potential energy. ToBY also failed on one of the thrown objects (stimulus #20), which paused when it reached its apex, and on one other object (stimulus #19) which had a failure in the trajectory construction phase.

Stimulus Category	Stimulus Number	Ground Truth	Human Judgment	ToBY Judgment	ToBY Expert
Static Objects	2	Inanimate	3%	Inanimate	Static Object
	16	Inanimate	6%	Inanimate	Static Object
Thrown Objects	7	Inanimate	44%	Inanimate	Energy
	13	Inanimate	*53%*	Inanimate	Energy
	20	Animate	78%	*Inanimate*	Straight Line
	25	Animate	81%	Animate	Energy
Circular Movements	5	Animate	59%	Animate	Energy
	8	Animate	81%	Animate	Energy
	17	Animate	81%	Animate	Straight Line
	26	Animate	78%	Animate	Acc. Sign Change
	29	Animate	56%	Animate	Energy
Straight Line Movements	4	Inanimate	47%	Inanimate	Straight Line
	11	Inanimate	36%	Inanimate	Straight Line
	22	Inanimate	30%	Inanimate	Straight Line
	27	Animate	53%	Animate	Energy
	15	Inanimate	37%	Inanimate	Straight Line
	24	Animate	75%	Animate	Energy
Pendula	3	Inanimate	16%	*Animate*	Energy
	10	Inanimate	12%	*Animate*	Acc. Sign Change
	21	Inanimate	31%	*Animate*	Acc. Sign Change
	30	Inanimate	19%	*Animate*	Acc. Sign Change
	12	Inanimate	6%	*Animate*	Acc. Sign Change
Erratic Movements	1	Animate	97%	Animate	Energy
	6	Animate	75%	Animate	Acc. Sign Change
	9	Animate	*31%*	Animate	Acc. Sign Change
	14	Animate	75%	Animate	Acc. Sign Change
	18	Animate	87%	Animate	Straight Line
	19	Animate	93%	*Inanimate*	Little Data
	23	Animate	81%	Animate	Energy
	28	Animate	90%	Animate	Straight Line

Figure 7-8: Comparison of human animacy judgments with judgments produced by ToBY for each of the stimuli from figure 7-6. Column 3 is the ground truth, that is, whether the trajectory actually came from an animate or inanimate source. Column 4 shows the percentage of human subjects who considered the stimulus to be animate. Column 5 shows the animacy judgment of ToBY, and column 6 shows the expert that contributed that decision. Italic items in the human or ToBY judgment columns indicate a disagreement with the ground truth.

7.4 Summary

The distinction between animate and inanimate is a fundamental classification that humans as young as 6 months readily perform. Based on observations that humans can perform these judgments using only spatio-temporal signatures, this chapter presented an implementation of a few naive rules for identifying animate objects. Using only the impoverished stimuli from the attentional system, and without any additional context, adults were quite capable of classifying animate and inanimate stimuli. While the set of experts explored in this chapter is certainly insufficient to capture all classes of stimuli, as the pendulum example illustrates, these five simple rules are sufficient to explain a relatively broad class of motion profiles. These simple algorithms (like the experts presented here) may provide a quick first step, but do not begin to make the same kinds of contextual judgments that humans use.

Chapter 8

Detecting Faces and Head Pose

> *An eye can threaten like a loaded and levelled gun, or can insult like hissing or kicking; or, in its altered mood, by beams of kindness, it can make the heart dance with joy.* – Emerson (1860)

Eye direction, body posture, and head orientation are all important social cues in human interactions. By observing these cues in another individual, people naturally make assumptions about their attentional state, attribute emotional states, and interpret behavioral goals and desires. These cues are never perfect predictors; a person can easily look in one direction but actually attend to something on the other side of the room. However, the orientation of the head, eyes, and body are part of the natural behavioral repertoire.

The ability to detect another creature looking at you is critical for many species. Many vertebrates, from snakes (Burghardt, 1990), to chickens (Ristau, 1991*a*), to primates (Povinelli & Preuss, 1995), have been observed to change their behavior based on whether or not eyes are gazing at them. In humans, eye contact serves a variety of social functions, from indicating interest to displaying aggression. Primates have further developed this ability to distinguish what another creature is looking at - that is, to follow and extrapolate its direction of gaze.

Gaze direction in particular is used for a wide variety of social signals (Nummenmaa, 1964). These cues are so integrated into our social behavior that it is difficult to imagine what it would be like without them. However, the importance of these social cues can be observed by considering a case of abnormal development. Individuals with autism do not maintain eye contact, or recognize pointing gestures, or understand simple social conventions. While their perceptual and abstract reasoning skills can be profound, autistics are severely disabled in our society.

Gaze direction can also be a critical element of social learning. Like a pointing gesture, gaze direction serves to indicate what object an individual is currently considering. While infants initially lack many social conventions (understanding pointing gestures may not occur until the end of the first year), recognition of eye contact is present from as early as the first month (Frith, 1990; Thayer, 1977). Detection of eye direction is believed to be a critical precursor of linguistic development (Scaife &

Bruner, 1975), theory of mind (Baron-Cohen, 1995), and social learning and scaffolding (Wood et al., 1976).

Computationally determining direction of gaze is a complex process that places very stringent requirements on a system. Detection of cues such as gaze direction and head orientation requires very high accuracy processing to be done quickly and reliably. People are extremely accurate in gauging the gaze direction of another individual, often being able to localize the target with only a single degree of difference (Nummenmaa, 1964) or even by accounting for minor variations in eye position caused by changes in vergence angle (Butterworth, 1991). Meaning is also conveyed by the dynamic aspects of gaze; a quick glance might last only a fraction of a second and yet carry a very significant social meaning. These challenges are compounded by difficulties imposed by building systems that can also generate these same social cues. Cameras that must move rapidly are more difficult to position accurately, and any moving camera also requires more complex image processing.

Because of the complexities involved in processing these cues, many research programs have focused on individual components of this task: detecting faces in an arbitrary visual scene, determining head orientation given a face location, or tracking gaze direction given a fixed location and orientation. Research on detecting faces in visual scenes has been the focus of numerous papers, conferences, and books (for example, the Automatic Face and Gesture Recognition conference sponsored by IEEE will celebrate its fifth year in 2002). Many of these research projects have focused on developing very accurate, but computationally expensive, techniques for finding faces (Rowley et al., 1995; Turk & Pentland, 1991; Sung & Poggio, 1994). A few more recent projects have attempted to build real-time detection systems (Darrell et al., 1998*a*; Jones & Viola, 2001). Other research has focused on the tracking of eyes and facial features for video conferencing (Graf et al., 1996; Maurer & von der Malsburg, 1996), as a user interface (Baluja & Pomerleau, 1994; Heinzmann & Zelinsky, 1997), or in animation (Terzopoulous & Waters, 1991); however, these techniques generally begin with calibrated high resolution images where the face dominates the visual field. Finally, a few systems have attempted to detect gaze direction given constrained visual scenes (Kozima, 1998) or by using an active infrared system that uses scleral reflections (the same reflections that cause red-eye in normal photography) (Morimoto et al., 1998).

For an active vision system in unstructured environments, these problems acquire a hierarchical structure; before finding gaze direction, the head location and orientation must be identified. These tasks have different sensory requirements. Detecting faces requires a wide field of view. Determining gaze direction and head orientation requires very high acuity in a localized region. The visual systems of Cog and Lazlo can accommodate some of these demands. The fields of view of the two cameras in each eye allow for both a wide field of view and a central region of high acuity. However, the resolution of these systems are insufficient for detecting gaze direction with a high degree of accuracy when a person is more than a few feet from the robot. To address this problem, we will rely upon head orientation rather than gaze direction as an indicator of attention. While this restriction does limit the behavior of the system, orientation is still a good measurement of attention and can be more easily determined by a computational process. With additional hardware in the form of a higher resolution camera, a lens with a tighter field of view, or with a computer-controlled zoom, information on gaze direction could be collected and used in the same ways that head orientation will be used.

To detect head orientation of a person within the robot's wide field of view, a five stage algorithm was developed:

1. Whenever a potential target is selected in the wide field of view by the attention system, the robot moves its eyes and head to foveate the object.

2. A skin color pre-filter is used to detect candidate face locations in the foveal image.

3. Two types of shape metrics are applied to candidate locations to verify the presence of a face.

4. A software zoom is used to extract the maximum detail from the location that contains a face.

5. Verified face locations are processed to find features including the eyes and mouth, which are then used to determine a rough estimate of head pose.

These five stages of processing are shown in figure 8-1. The dynamic nature of the task requires that all of the stages of processing happen in real-time and with minimal latencies. In building these algorithms, these temporal requirements will be more critical than the accuracy requirements; missing a face in one particular image is not as important when you have another chance to detect in the very next frame.

The following five sections will describe each of these processing stages in detail. Section 8.6 will return to the problem of acquiring images of eyes in sufficient detail when there is a person within a few feet of the robot.

8.1 Foveating the Target

The first stage in detecting head pose focuses on bringing the face within the field of view of the foveal camera. This step accomplishes two goals: it allows the face to be imaged in sufficient detail for further processing and it also gives the human under observation the social cue that the robot is looking at them. Both of these aspects are critical for simplifying the processing that the robot performs. By bringing the face into the foveal field of view, a high acuity image can be obtained in which the face is likely to be the largest single object. This movement also serves as a social cue to the human, who often will respond in a way that makes the processing task of the robot simpler. When the robot turns to look at a person, it is common for that person to also turn toward the robot. This mutual orientation makes the robot's computational task easier by providing a motion stimulus and by aligning the face toward the robot. The person often will move so that the robot is more clearly visible, removing possible occlusions.

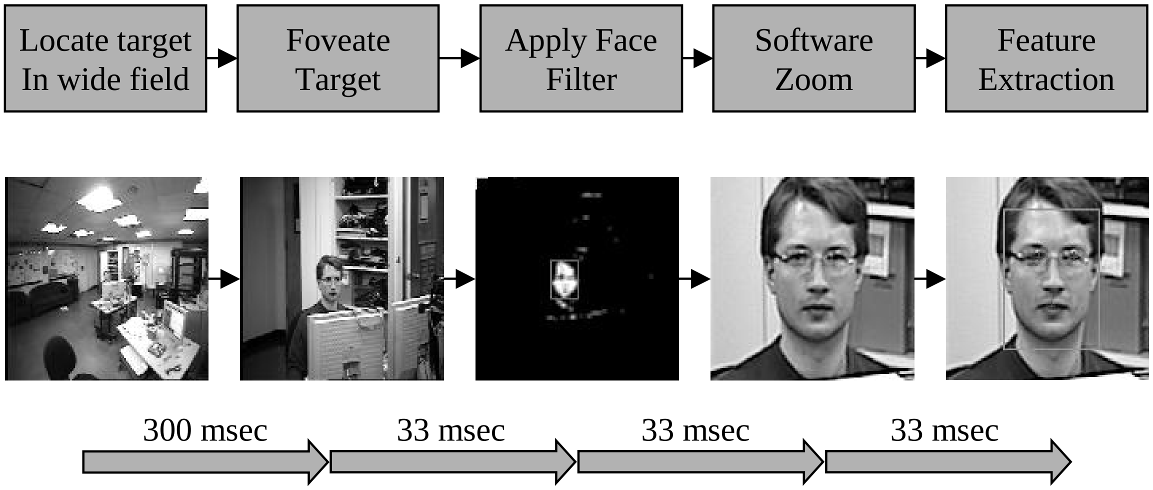


Figure 8-1: Five stages of processing for finding head pose. The attention system locates regions of interest in the peripheral image. These targets are brought into the field of view of the foveal image by a saccadic eye movement. Within the foveal image, a skin color prefilter locates candidate face locations which are then verified by a set of shape metrics. Verified faces result in a software zoom within the foveal image to that location. These faces are then processed to find feature locations (such as eyes and mouth) that can be used to determine head pose.

Possible targets are selected by the attention system using the same saliency criteria discussed in chapter 6. Recall that it is easy to bias the robot toward faces by increasing the attentional weighting for moving and skin-colored objects. The robot will then move its eyes and neck to look at and orient toward the object of interest. This sensorimotor skill is acquired through a reinforcement learning technique (see section 5.1.1). This action brings the target to the center of the peripheral camera, which is roughly aligned to be within the field of view of the foveal camera. However, the centers of the two fields are not necessarily the same due to parallax and misalignment of the camera mountings. To relate positions in the peripheral image to positions in the foveal image, a second sensorimotor mapping was learned. This mapping was acquired in two stages:

1. An estimate of the relative difference in scale between the fovea and peripheral cameras was obtained by observing the relative rates of optic flow in the two images while the eye was moving.

2. The foveal image was reduced by that scaling factor and used as a correlation mask to compare against the peripheral image. The best location match gives the location of the foveal center within the peripheral image

These two steps are sufficient to map locations in the foveal image to locations in the peripheral image.

The difference in scale between the two images is obtained by measuring differences in optic flow. To measure these rates of change, one option is to keep the eyes in a fixed position while observing a moving object. If the object is tracked in both images, then the difference in the rate of movement is the same as the difference in scale. However, this approach has difficulties both in isolating the target and in the accuracy of the scale measurement. Because it is rarely the case that there is only a single source of movement from a rigid object in the world, it would be necessary to verify that the same object was being observed in both images. While this might be accomplished by looking at some featural properties of the object (other than size) such as color content, texture, or shape, these features may be difficult to compare across images without additional camera calibration as the color content, exposure, and background content will vary between the images. Further, the accuracy of the velocity measurement would be limited by the resolution of the camera images. Objects moving at speeds that are easy to measure accurately in the peripheral image may move too quickly in the foveal image for accurate tracking. An alternative approach is to wait until there is no motion in the visual scene and then to move the cameras themselves. This self-generated movement results in perceived motion of the entire background. The two cameras in each eye are fixed with respect to each other; as the eye moves, the two cameras remain in the same relative positions to each other. Because of this mechanical arrangement, as the eyes move the rate at which objects move across the visual fields gives the difference in scale between the images. This background motion is measured using standard correlation-based optic flow algorithms (Horn, 1986). This alternative technique solves both the problems of single object tracking. No single object need be selected, so there is no associated matching problem. Also, the eyes can be moved at a variety of speeds to find a speed at which the optic flow can be reliably measured in both images.

Once the difference in scale has been localized, it is possible to complete the mapping of foveal image points to peripheral image points by establishing the location of the foveal center within the peripheral image. This mapping assumes that there is relatively little difference in rotation about the image axis and that the scale across the image plane is relatively constant. The rotation constraint basically imposes that "up" is the same for both image planes. This is a reasonable assumption for the camera systems on Cog and Lazlo, as the cameras have a mechanical alignment tab that is relatively easy to position. The assumption of a constant scale can be observed both from the qualitative observation that the entire foveal image and the center portion of the peripheral image (see figure 4-2) and quantitatively in the linearity of the saccade mapping for image positions near the center of the peripheral image. Given these assumptions, the location of the relative field centers can be identified by reducing the size of the foveal image by the ratio of the scale factors between images, in essence, by making the foveal image to be the same scale as the peripheral image. This reduced image is then used as a correlation mask to find the best match within the peripheral image plane. The location of the best match gives the location of the foveal image within the periphery. Using this information, any image point in the foveal image can be mapped to the peripheral image by considering the vector from the foveal image center, scaling that vector by the ratio of optic flow scales, and shifting the vector to the center point of the correlation best match in the peripheral image. Similarly, peripheral image points can be mapped to foveal image points, although obviously not every peripheral image position can be mapped to a foveal position.

Both of these steps can be done automatically without any need for human intervention. Further, these steps rely on the active nature of the vision system in acquiring the scale difference between cameras. When a target is foveated by the attention system, this mapping can be used to identify potential regions in the foveal image and to bias the verification stages in the foveal image. In practice, the ratio between scale factors between the two images is determined by the optical characteristics of the camera and lens and is independent of mounting. This ratio was found to be 4.0 for both eyes on both Cog and Lazlo. The offset positions of the best correlation match do depend on the mounting and thus vary between each eye and can easily change whenever the eyes are re-assembled. The center of the fovea image tends to be very close to the center of the image in each case. For example, at one measurement for Cog's right eye, the center of the reduced foveal image in the periphery was at pixel 268 (of 512) horizontally and 248 (of 512) vertically. The correlation measurement can easily be re-evaluated whenever maintenance is done on the robot.

8.2 Skin Color Filtering

Once a target has been brought within the foveal field of view, the robot must determine whether it is looking at a face or some other object. The first step in this analysis is to use a skin-color filter to remove background effects and as a first pass to remove non-face objects. The skin color filter used on the foveal image is the same as the one used by the attention system for determining saliency that was described in section 6.3.3. The skin color filter used directly in the attention system produces a saliency image, that is, for each pixel in the image, the filter produces a rating from 0 to 1 on how well the pixel matches the skin tone region of color space. For the attention saliency map, this number is scaled to 0–255 in order to match the other filter outputs. For filtering face images, the unscaled value is thresholded at 0.50. Pixels that pass this threshold are retained in the post-filtering image. Pixels that fail the threshold are set to black in the post-filtering image. (The next stage of processing will not depend on the color content, only on the intensity. By setting other pixels to black, they are effectively removed from the processing stream.) Regions are identified in the skin-filtered image using a four-neighbor connected components algorithm identical to the one used by the attention system. The result of this processing gives a set of connected regions for each image that contain skin-colored pixels.

The skin color filter is not an ideal filter for detecting faces. First, some items that are skin but are not faces will also pass through the filter. Second, the skin filter accepts certain objects that are not skin. Wooden doors, cardboard, and some pressed-board notebooks are accepted by the skin color filter. These objects are common in most office environments, including the laboratory spaces that our robots occupy. To combat this, a second verification step must be provided to determine whether the observed object is actually a face.

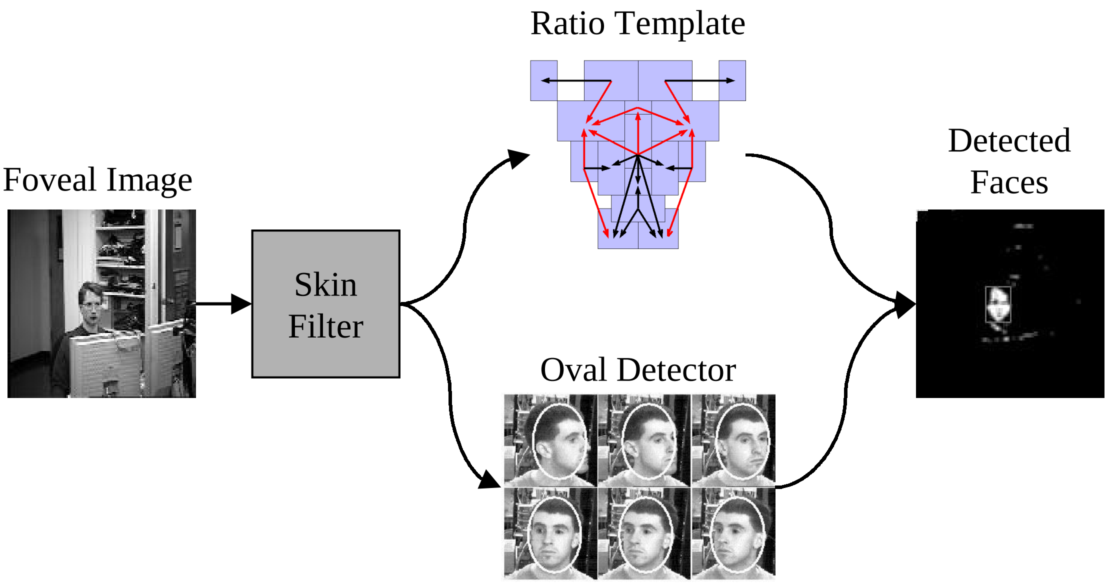


Figure 8-2: The face detection system. Once an object is brought within the foveal field of view, two stages of processing are applied to determine whether or not a face is present. First, a skin-color filter is applied to determine candidate areas. Second, a pair of shape metrics are used to evaluate whether the object has the proper shape. If both agree, then the object is classified as a face.

8.3 Detecting Faces

The choice of a face detection algorithm was based on three criteria. First, it must be a relatively simple computation that can be performed in real time. Second, the technique must perform well under social conditions, that is, in an unstructured environment where people are most likely to be looking directly at the robot. Third, it should be a biologically plausible technique. To verify that a location is actually a face, two algorithms based on shape are used (see figure 8-2). The first verifies that the exterior shape (the boundary of the object) is roughly oval and the second verifies that the internal structure resembles a face (it has eyes, a mouth, etc.). These metrics are made computationally tractable in real-time by applying the metric not to the complete image at all scales but rather only to the connected components of the skin-filtered image at scales that match the size of the connected component. Both of these techniques were designed to detect frontal views of faces, which matches the social constraint. Finally, these metrics are rough estimates of the kind of processing that is believed that infants are capable of performing (Fagan, 1976). Infants are sensitive not only to the exterior shape of an object but also to the rough arrangement of features within that object.

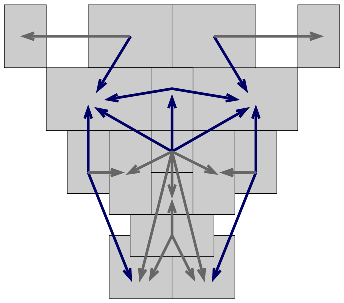


Figure 8-3: A ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows). Darker arrows are statistically more important in making the classification and are computed first to allow real-time rates.

8.3.1 Exterior Shape Metric: Oval Detector

The exterior metric attempts to verify that the boundary of the object is roughly oval. The algorithm for this verification was implemented and applied to face detection by Artur Arsenio, Jessica Banks, and Paul Fitzpatrick within our laboratory. Given a connected component from the skin-color filter, the oval detector attempts to find edge boundaries that roughly match an oval of given proportions. In applying the detector to the skin-color patches, two possible center points are considered: the centroid of the patch and the center of the bounding box. These two variants allow for slight variations in the exterior boundary to have little impact on further processing. Given these center locations, the size of the oval is fixed to be the maximum size of the bounding box, with the caveat that the oval maintain an aspect ratio that favors the vertical dimension.

Given a center position and a size, the algorithm searches through a set of aspect ratios that could be contained within that bounding box. For each possible oval, the algorithm looks for edges in the image at 32 points along the upper 300 degrees of the oval (allowing the lower sixty degrees to vary based on the neck presence or absence). If 80% or more of these points have an image gradient that flows in toward the center of the oval, then the location is considered to be a successful oval.

This exterior metric has been successfully used on its own to detect faces within unstructured visual scenes by Arsenio and Banks. The detector is used in this work primarily to remove the presence of objects that pass the skin-color filter but that are completely inappropriate (such as doors and notebooks). The more critical judgment in distinguishing hands and other objects with complex structure from faces is made by evaluating the interior shape metric.

8.3.2 Interior Shape Metric: Ratio Template

To evaluate the internal shape of the target object, a template-based algorithm called ratio templates (Sinha, 1994) was selected. The ratio template algorithm was designed to detect frontal views of faces under varying lighting conditions, and is an extension of classical template approaches (Sinha, 1996). However, rather than being based on particular pixel values as a normal template would, the ratio template is based on the gradients between template regions. For detecting faces, the ratio template attempts to capitalize on lighting-invariant features of faces that are based upon the facial structure of normal individuals. For example, the eye sockets tend to be recessed and thus often appear darker than the surrounding face. The ratio template uses these regularities to detect faces. While other techniques handle rotational invariants more accurately (Sung & Poggio, 1994), the simplicity of the ratio template algorithm allows us to operate in real time while detecting faces that are most likely to be engaged in social interactions. Ratio templates also offer multiple levels of biological plausibility; templates can be either hand-coded or learned adaptively from qualitative image invariants (Sinha, 1994).

A ratio template is composed of a number of regions and a number of relations, as shown in Figure 8-3. For each target location in the image, a template comparison is performed using a special set of comparison rules. First, the template is scaled to match the bounding box of the connected component patch identified from the skin-color filter. The template is overlayed on a grayscale copy of the filtered image at the location of the connected component. Each region in the template is convolved with the grayscale image to give the average grayscale value for that region. Relations are comparisons between region values, for example, between the "left forehead" region and the "left temple" region. The relation is satisfied if the ratio of the first region to the second region exceeds a constant value (in our case, 1.1). This ratio allows us to compare the intensities of regions without relying on the absolute intensity of an area. In figure 8-3, each arrow indicates a relation, with the head of the arrow denoting the second region (the denominator of the ratio). We have adapted the standard ratio template algorithm to process video streams (Scassellati, 1998*b*). In doing so, we require the absolute difference between the regions to exceed a noise threshold, in order to eliminate false positive responses for small, noisy grayscale values. In practice, for each connected component in the skin-color filtered image, a small number of possible face locations are analyzed by the ratio template region. A small range of motion of the center and a small range of scales are allowed in the comparison, with neither exceeding 10% of the original values.

Improving the Speed of Ratio Templates

To improve the speed of the ratio template algorithm, we have implemented an early-abort optimization. At the suggestion of Sinha (1997), we further classified the relations of our ratio-template into two categories: eleven essential relations, shown as black arrows in figure 8-3, and twelve confirming relations, shown as gray arrows. We performed a post-hoc analysis of this division upon approximately ten minutes of video feed in which one of three subjects was always in view. For this post-hoc analysis, an arbitrary threshold of eighteen of the twenty-three relations was required to be classified as a face. This threshold eliminated virtually all false positive detections while retaining at least one detected face in each image. An analysis of the detected faces indicated that at least ten of the eleven essential relations were always satis-

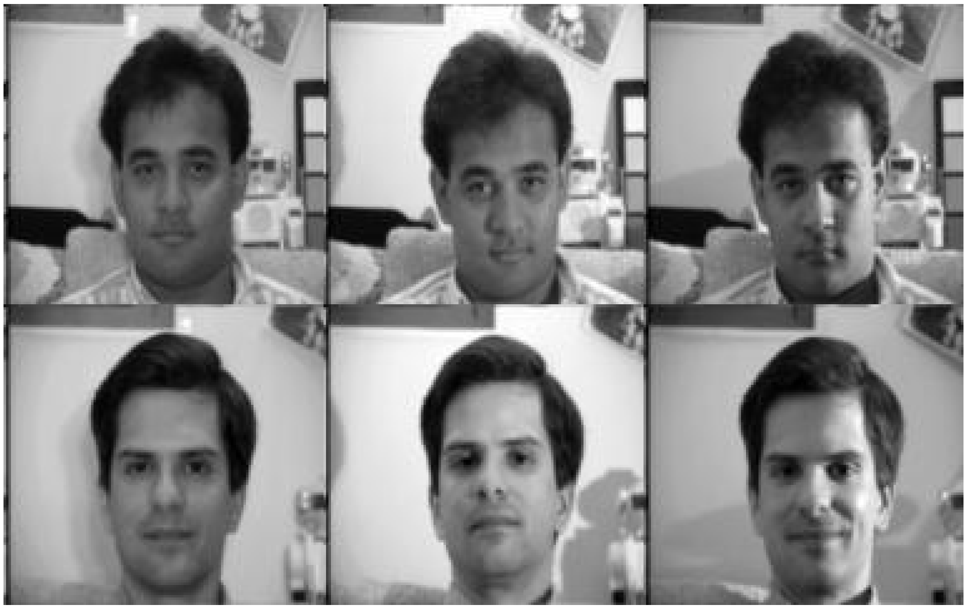


Figure 8-4: Six of the static test images from Turk & Pentland (1991) used to evaluate the ratio template face detector. Each face appears in the test set with three lighting conditions, head-on (left), from 45 degrees (center), and from 90 degrees (right). The ratio template correctly detected 71% of the faces in the database, including each of these faces except for the center image from the first row.

fied. None of the confirming relations achieved that level of specificity. Based on this analysis, we established a new set of thresholds for face detection: ten of the eleven essential relations and eight of the twelve confirming relations must be satisfied. As soon as two or more of the essential relations have failed, we can reject the location as a face. This optimization allows for very rapid rejection of non-face patches and increases the overall speed of the ratio template algorithm by a factor of four.

Static Evaluation of Ratio Templates

To evaluate the static performance of the ratio template algorithm, we ran the algorithm on a test set of static face images first used by Turk & Pentland (1991). The database contains images for 16 subjects, each photographed under three different lighting conditions and three different head rotations.

To test lighting invariance, we considered only the images with an upright head position at a single scale, giving a test set of 48 images under lighting conditions with the primary light source at 90 degrees, 45 degrees, and head-on. Figure 8-4 shows the images from two of the subjects under each lighting condition. The ratio template algorithm detected 34 of the 48 test faces. Of the 14 faces that were missed, nine were the result of three subjects that failed to be detected under any lighting conditions. One of these subjects had a full beard, while another had very dark rimmed glasses, both of which seem to be handled poorly by the static detection algorithm. Of the remaining five misses, two were from the 90 degree lighting condition, two from the 45 degree lighting condition, and one from the head-on condition. While this detection rate (71%) is considerably lower than other face detection schemes (Rowley et al., 1995; Turk & Pentland, 1991; Sung & Poggio, 1994), this result is a poor indicator of the performance of the algorithm in a complete, behaving system (see section 8.6.1 below).

Using the real-time system, we determined approximate rotational ranges of the ratio template algorithm. Subjects began looking directly at the camera and then rotated their head until the system failed to detect a face. Across ten subjects, the average ranges were ±30 degrees pitch, ±30 degrees yaw, and ±20 degrees roll.

8.4 Software Zoom

If a skin-colored patch is accepted by both the interior and exterior shape metric, the patch is declared to be a face. If multiple face locations are present in a single foveal image, the larger face (which is almost always the face closer to the camera) is selected as the primary target. This selection process is arbitrary, but it does allow the robot to have reasonable behavior when faced with large crowds (or tour groups). Note also that this choice does not drive the outward signs of the robot's attentional state; the eye and head motion is still controlled directly by the attention system. The net effect of this decision is to make people closer to the robot more important than people further away.

Once a face is selected, the software controlling the frame grabber on the foveal image is reconfigured to capture an image of the face area at the maximal resolution. To allow face verification and feature extraction to occur in parallel, Cog employs two different hardware frame grabbers for each foveal camera. The first of these always captures images of the entire field of view at the fixed resolution used for image processing (128 by 128), while the other is used as a "software zoom" to capture a sub-region of the complete field of view at this same resolution. The field of view and center position of this zoomed image is set in software to maximize the information coming from the frame grabber. Ideally, the zoomed image is a 128 by 128 subset of the full 640 by 480 NTSC camera signal centered on the face. A face at approximately three meters from the robot will fill this zoomed image at maximum resolution. Faces further away will be smaller in the zoomed image, as there is no additional resolution that can be obtained from the statically-configured cameras. Faces closer to the robot can be fit exactly to this size.

It is important to note that because the grabber parameters are controlled directly, the location of the face in the zoomed image (or of any known point in the full foveal image) can easily be computed. This property will allow us to begin to analyze the face within the zoomed image immediately without any further computation. The only differences will be induced by the delay between acquiring and processing the full foveal image and the response time for acquiring a new zoomed image. In general, the processing of the face finding algorithm runs at real-time rates (30 Hz), so the delay between the original acquisition and the availability of the zoomed image is the time to process the full frame (1/30 of a second) plus the time to change the grabber

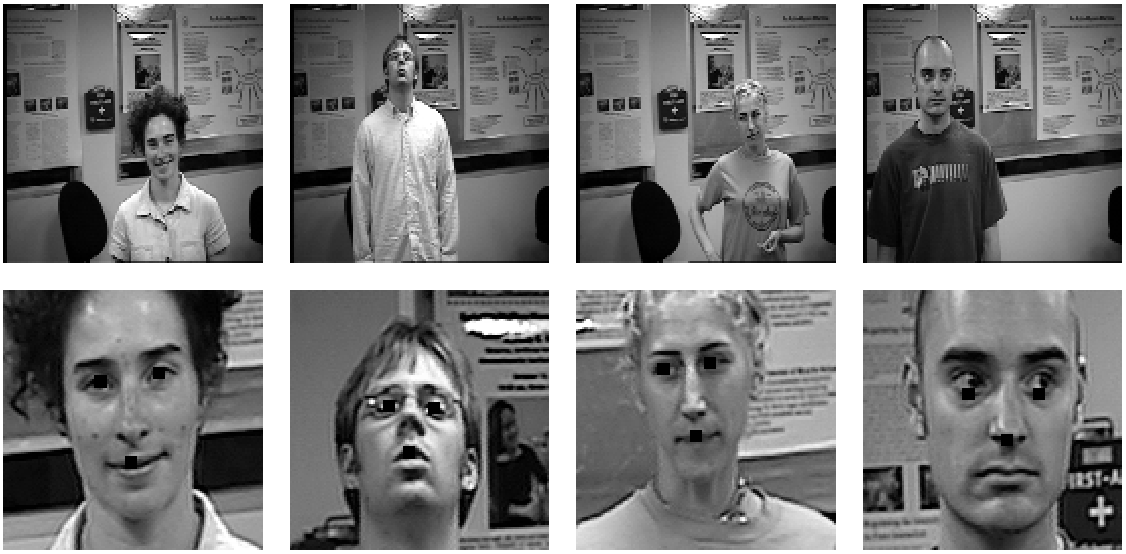


Figure 8-5: Eye and mouth regions detected by the multi-stage classifier system from figure 8-1. The top row contains the foveal images acquired immediately after a saccade, while the lower row contains the result of the feature identification system. The positions of the eyes and the mouth are shown with black dots. The first three cases were successful, but the last examples failed due to a mis-match of the feature set to the nostrils rather than the mouth.

parameters and acquire a new image (between 1/30 and 1/15 of a second).

8.5 Finding Facial Features to Estimate Head Pose

Once a face has been located and a zoomed image of the face has been acquired, the final step is to locate features within the face which can be used to interpret head pose. The approach used here is similar to the work of Heinzmann & Zelinsky (1997) as well as Loy et al. (2000), but lacks some of the sophistication used in either of these systems. Heinzmann and Zelinsky use Kalman filters to track triplets of corner features with one triplet for each eye and another triplet for the mouth, while Loy, Holden, and Owens use seven features based on the corners of the mouth, the outside corners of the eyes, the center of the upper and lower lips, and the midpoint between the nostrils. Both of these systems, and the majority like them, are tracking systems that rely upon a manual or semi-automated initialization phase in which a user either selects the initial locations of target features or in which the observed individual must perform some fixed action (such as orienting to known positions in the environment). The simplified system presented here has neither the accuracy nor the computational generality that other techniques possess, but it does have the necessary autonomy that these robotics applications require.

To determine a rough estimate of head pose, three salient facial features are detected in each frame: the centers of each eye, and the center of the mouth. These three features, combined with the bounding box of the face area, are sufficient to determine the orientation of the head (Gee & Cipolla, 1994). Each of these features can be defined by a low-intensity patch in a specific region of the face template. The algorithm presented here uses an iterative refinement technique to localize these region centers that is similar to simulated annealing (Kirpatrick et al., 1993). The algorithm operates as follows:

1. Localize the mouth region by searching within a limited range of the face template for the centroid of a large, low-intensity patch in the skin-filtered image.

2. Localize the eye regions by searching for a pair of low-intensity patches that are symmetric about the line parallel to the principle axis of skin-colored patch which passes through the mouth center.

3. Extrapolate head pose based on these three points.

Although we allow for an iterative refinement stage in the first two steps of this algorithm, the complete process must operate at real-time rates.

The mouth region is identified as the largest patch of low intensity that occurs within the lower 60–80% of the face region. The mouth is identified by starting with a relatively high skin-filter threshold and annealing the threshold rate by gradually reducing it and allowing the center point of the detected region to deviate slightly at each time step. In this way, the center point initially is placed in a relatively large region that may include pixels that are actually skin, but that as the threshold is reduced, more skin pixels are correctly classified and correspondingly more non-skin pixels are incorrectly classified. The skin threshold is dropped logarithmically until either the center point of the detected mouth region remains stationary for three consecutive iterations or until ten iterations have passed. This process allows the center point to deviate from its original position to more accurately reflect the midline of the face while maintaining a relatively quick and simple algorithm.

Once the mouth region has been localized, a similar annealing step is performed for the two eyes. This process has one additional complication in that the eye regions are simultaneously searching for two intensity minima while at the same time attempting to maintain a symmetry requirement. A line of symmetry for the face is determined by considering all lines parallel to the principle axis of the oval (the first moment of the pixels passing threshold) and selecting the line that passes through the point selected as the mouth center. The search range for the eyes occurs within the upper 30–50% of the face region. At each time step in the annealing process, the skin filter threshold is reduced by a logarithmic scale and the center point of the two eye regions are allowed to move toward the new centers of low-intensity regions with the constraint that the two points must move to maintain the same symmetry relationship.

Figure 8-5 shows the result of this algorithm on a few samples images from the full foveal field of view (top) and the zoomed image (bottom). The locations of the eyes and mouth are shown as black dots. The system successfully located the mouth and eye regions of the first three subjects shown here, but had some failures on the final subject. The failure shown here, which was a typical failure mode for the complete system, had a match for the mouth point that was still on the main symmetry axis, but was located more toward the nose than the mouth. This failure had relatively little impact on the detection of head orientation, as the symmetry axis was unaffected. Another more serious failure was the confusion of the eye regions with the eyebrow regions. The simulated annealing techniques were usually sufficient for guiding selections toward the larger, darker, eye regions, but there were a few instances seen in testing in which the eye search was caught by the local minima of the eyebrows.

A triangle defined by these points in relation to the bounding box of the entire face can be used to determine head pose (Gee & Cipolla, 1994). The relative positions of these points within the bounding box provides a quick estimate of the yaw and pitch of the head, while the rotation of the skin region and the rotation of the symmetry axis gives the roll of the head. The head posture system was evaluated on ten subjects to determine the accuracy of each head posture. Subjects were asked to stand anywhere in the room between 2 and 6 meters from the robot. Once the robot had oriented toward them, they were asked to move their heads left to right (yaw), then up and down (tilt), and finally to tilt their head from shoulder to shoulder (roll). At distances of up to six meters, the yaw position could be determined with an accuracy of between five and ten degrees, while the other two directions (tilt and roll) could be localized within ten and fifteen degrees. Chapter 10 will make use of this information to provide the robot with behaviors that allow it to operate on objects based on the attentional state (represented here by orientation) of an instructor.

8.6 Extracting Eye Images

While the system presented in this chapter concentrates on extracting the basic featural properties of head orientation, the ideal system would also be responsive to gaze direction. Although the perceptual problems of detecting gaze are not addressed here, we have demonstrated that at very close distances, the robot can obtain high-resolution images of the eyes using the behaviors outlined above. With a person standing less than four feet in front of the robot, the location of the eyes from the ratio template (or from the eye locations derived in the previous section) can be used to extract a high-resolution zoomed image that contains only the eye. Figure 8-6 shows a few examples of the foveal images and the resulting zoomed image.

8.6.1 Dynamic Evaluation of Eye Finding

The evaluation of this system must be based on the behavior that it produces, which can often be difficult to quantify. The system succeeds when it eventually finds a face and is able to extract a high resolution image of an eye. However, to compare the performance of the entire system with the performance of the ratio template algorithm on static images, a strawman quantitative analysis of a single behavior was studied. Subjects were asked to sit within 4 feet of Lazlo. The subject was to

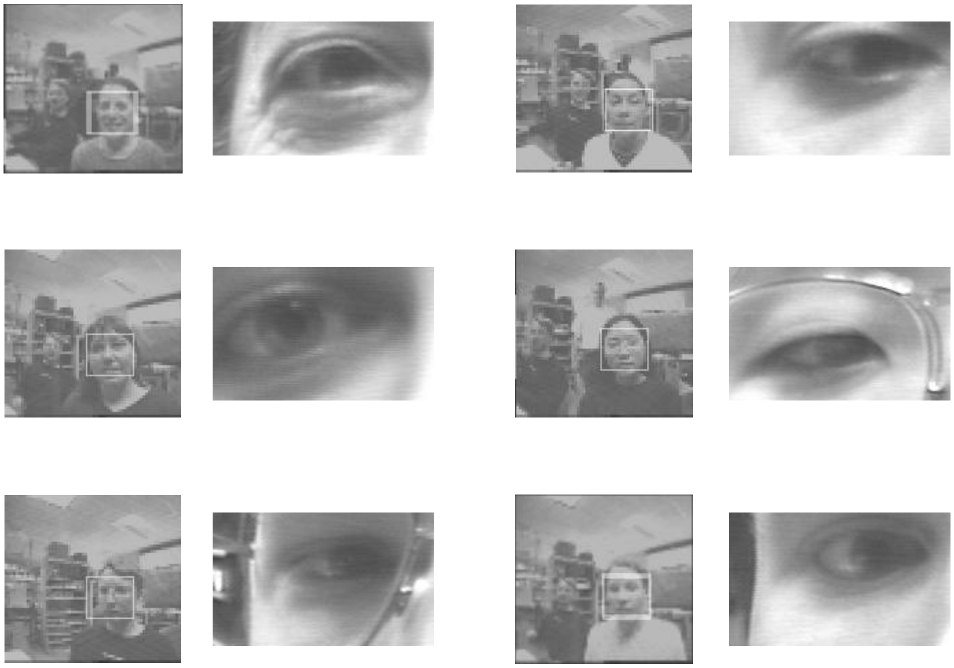


Figure 8-6: A selection of faces and eyes identified by the robot. Potential faces are located in the peripheral image. The robot then saccades to the target to obtain a high-resolution image of the eye from the narrow field-of-view camera.

remain stationary during each trial, but was encouraged to move to different locations between trials. These tests were conducted in the complex, cluttered background of our laboratory workspace (similar to the images shown in figure 8-1).

For each behavioral trial, the system began with the eyes in a fixed position, roughly centered in the visual field. The system was allowed one saccade to foveate the subject's right eye (an arbitrary choice). The system used the skin color filtering and ratio template face detection routines to generate a stream of potential face locations. Once a face was detected, and remained stable (within an error threshold) for six cycles (indicating the person had remained stationary), the system attempted to saccade to that location and extract the zoomed eye image. 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.

This quantitative analysis of the system is extremely promising. However, the true test of the behavioral system is in eventually obtaining the goal. Even in this simple analysis, we can begin to see that the total behavior of the system may be able to correct for errors in individual components of the system. For example, one incorrect face identification was a temporary effect between part of the subject's clothing and the background. Once the system had shifted its gaze to the (false) face location, the location no longer appeared face-like. Without the arbitrary imposition of behavioral trials, the natural behavior of the system would then have been to saccade to what it did consider a face, achieving the original goal.

If our behavioral test had allowed for a second chance to obtain the goal, the failure rate can be estimated as the product of the failure rates for each individual trial. If we assume that these are independent saccades, the probability of failure for a two-attempt behavior becomes $0.06 \times 0.06 = .0036$. As we allow for more and more corrective behavior, the stability of the system increases. While individual trials are probably not completely statistically independent, we can see from this example how the behavior of the system can be self-stabilizing without requiring extremely accurate perceptual tools.

Issues like these make quantitative analysis of behaving systems difficult and often misleading (Brooks, 1991*a*). Our system does not require a completely general-purpose gaze tracking engine. In a real-world environment, the humans to whom the robot must attend in order to gain the benefits of social interaction are generally cooperative. They are attempting to be seen by the robot, keeping their own attention focused on the robot, facing toward it, and often unconsciously moving to try to attract its attention. Further, the system need not be completely accurate on every timestep; its behavior need only converge to the correct solution. If the system can adequately recognize these situations, then it has fulfilled its purpose.

8.7 Summary

Faces have a special status in human visual perception. Social interactions in particular are critically dependent on being able to locate people in the visual scene, and detecting faces is an obvious partial solution to that requirement. Further, the focus of attention of an individual is commonly reflected in the individual's posture, head orientation, and gaze direction.

We have implemented a system that identifies faces in a visual scene, attempts to obtain an image of the face in as much resolution as possible, and then calculates head pose based on the locations of certain facial features (the eyes and mouth). We have also extended this system on Lazlo to find eyes in the visual scene, and to obtain high resolution images for future processing. We will return to the use of head pose as an indicator of attention and as a trigger for joint reference behavior in chapter 10. However, before we can build on these foundational theory of mind components, chapter 9 will introduce a basic social learning behavior which will have a vital role in demonstrating the implications of theory of mind skills.

Chapter 9

A Simple Mechanism for Social Learning

Imitation is to understanding people as physical manipulation is to understanding things. – Meltzoff & Moore (1994, p. 83)

Social learning skills provide a powerful mechanism for an observer to acquire behaviors and knowledge from a skilled individual (the model). Biologists have delineated many different forms of social learning, from direct tutelage in which the model actively attempts to instruct the novice to social facilitation in which merely the contingent presence of the model and the objects involved in the action provide the necessary cues for learning (Galef, 1988; Whiten & Ham, 1992; Hauser, 1996). In particular, imitation is an extremely powerful mechanism for social learning which has received a great deal of interest from researchers in the fields of animal behavior and child development (for a review, see Dautenhahn & Nehaniv, 2001).

Similarly, social interaction can be a powerful way for transferring important skills, tasks, and information to a robot. The grand challenge, and the great hope, of imitation research in robotics is that this will provide a way for the robot to acquire new skills and actions on its own without additional programming. Imitation has been an active area of research within the robotics and artificial intelligence communities as a potential mechanism for overcoming the problems involved in building incrementally complex systems without expending enormous amounts of human effort. This issue, often called the scaling problem, is at the core of the motivation for many machine learning systems and has been examined from the perspective of behavior-based robotics by Tsotsos (1995). Schaal (1999) surveyed imitative learning specifically for humanoid systems and concluded that imitation was a rich problem space for the integration of perceptual systems, motor systems, and cognitive modeling.

Social cues provide the perceptual bias that most machine learning systems in complex environments crave. Many machine learning systems in robotics research operate in environments in which there is a high penalty for failure (such as falling off a cliff), a high level of environmental complexity (rough terrain or complex dynamic scenes), and unreliable or uncertain feedback mechanisms for determining whether an action was successful or not. Social cues such as a brief head nod, an encouraging tone of voice, or a pointing gesture provide structure to the environment by providing a mechanism for feedback. These social cues also serve to simplify the computational requirements of the robot by pointing out the aspects of a scene that are most relevant. For example, when demonstrating a complex process, the gaze direction of the instructor gives a good indication of the aspect of the task that is most attracting their attention and thus should be the object of attention of the observer. These simplifications allow for relatively simple machine learning techniques to be used without resorting to more complex statistical methods for extracting regularities from the environment.

9.1 Definitions

Great debates in animal cognition have focused on defining labels for different behaviors that allow social learning. Galef (1988), Whiten & Ham (1992), and Hauser (1996) have all made serious attempts at building taxonomies for describing different classes of social learning. The goal behind these descriptive attempts (and much of the work on imitation in animals) is to identify the minimal sets of cognitive abilities that underlie a particular observed behavior without attributing any additional competencies.[1] For our purposes, it is not necessary to attend to all of the finer distinctions between observational conditioning, goal emulation, social facilitation, imitation and the many other classes of social learning that have been proposed. However, one distinction used by Byrne (1999) is critically important: one type of imitation copies the organizational structure of a behavior while the other copies the surface form of the behavior. For example, suppose a robot was to observe a person picking up a paintbrush and applying paint to a wall. The robot could imitate the surface form of this event by moving its arm through a similar trajectory, perhaps even encountering a wall or a brush along the way. However, the underlying organizational structure of applying paint to a wall involves recognizing the intent of the action as well as the usefulness of the tool in accomplishing the goal. Note that this is the same distinction that was made in the studies of Meltzoff (1995) in which children as young as 18 months of age were able to repeat not only the surface form the behavior, but also could recognize and tended to respond with the underlying intended action.

In this work, we will use the word *imitate* to imply that the observer is not merely replicating the actions of the model but rather is attempting to achieve the goal of the model's action by performing a novel action similar to that observed in the model. The simpler action of copying the outward surface form of a movement will be called *mimicry*.[2] While the grand challenge is to build robots that *imitate*, the issues

[1] It is interesting to note that, at their core, each of these descriptions is really an attempt by human observers to suppress the attributions that their own theory of mind provides.

[2] Note that this is not a normally accepted definition. In most biological literature, the word *mimicry* indicates that the outward appearance of an animal has evolved to appear to be something

involved in building a robotic system that can *mimic* are still extremely challenging. Simpler mechanisms such as stimulus enhancement, emulation, and mimicry must address challenges such as determining what actions are relevant in the scene and finding conspecifics, while other challenges (such as determining the goal behind an action) are specific to this definition of imitation. It is an open question as to whether or not inferring intent is necessary to explain particular behaviors (Byrne, 1999). However, for a robot to fulfill the expectations of a human instructor, the robot must have a deeper understanding of the goal and intent of the task it is learning to perform.

In the next section, we will review the current research on robotic systems that imitate or mimic. Section 9.3 outlines four hard problems in building robots that imitate people and section 9.4 discuss how the social cues that humans naturally and intuitively provide could be used by a robot to solve these difficult problems. Much of the material in these two sections is drawn from a review by Breazeal & Scassellati (2001*b*). Section 9.5 introduces a mechanism based on the perceptual and motor skills outlined in previous chapters that allow an upper-torso robot to mimic socially presented stimuli.

9.2 Existing Studies of Imitation in Robotics

There are many, many perceptual, motor, and cognitive skills that are necessary to begin to address the specific problems of imitation. Figure 9-1 shows a small subset of the necessary behaviors which have been implemented or are currently under development by the Humanoid Robotics Group at MIT. Most of the listed skills represent the work of large communities of researchers, with individual books, journals, and conferences dedicated to each. The integration of each of these components is also a challenging topic by itself. For example, representing the dynamic interaction between different behaviors or understanding the compromises involved in using many different perceptual filters presents new sets of challenges.

To begin to address the specific problems of imitation, each robotics research team must make some simplifying assumptions and trade-offs. Simplifications in the hardware design, the computational architecture, the perceptual systems, the behavioral repertoire, and cognitive abilities allow a research team to address the more complex issues without implementing complete solutions to other problems. Each research team must be very careful to describe the assumptions that are made and the potential implications of these assumptions on the generality of their results. While these simplifications are at one level unavoidable, it is important to keep the big picture in mind.

Initial studies of social learning in robotics focused on allowing one robot to follow a second robot using simple perception (proximity and infrared sensors) through mazes (Hayes & Demiris, 1994) or an unknown landscape (Dautenhahn, 1995). Other

else, such as the eye spots common in the wing patterns of many butterflies (Hauser, 1996). While this overlap is unfortunate, other proposed labels are excessively long or awkward.

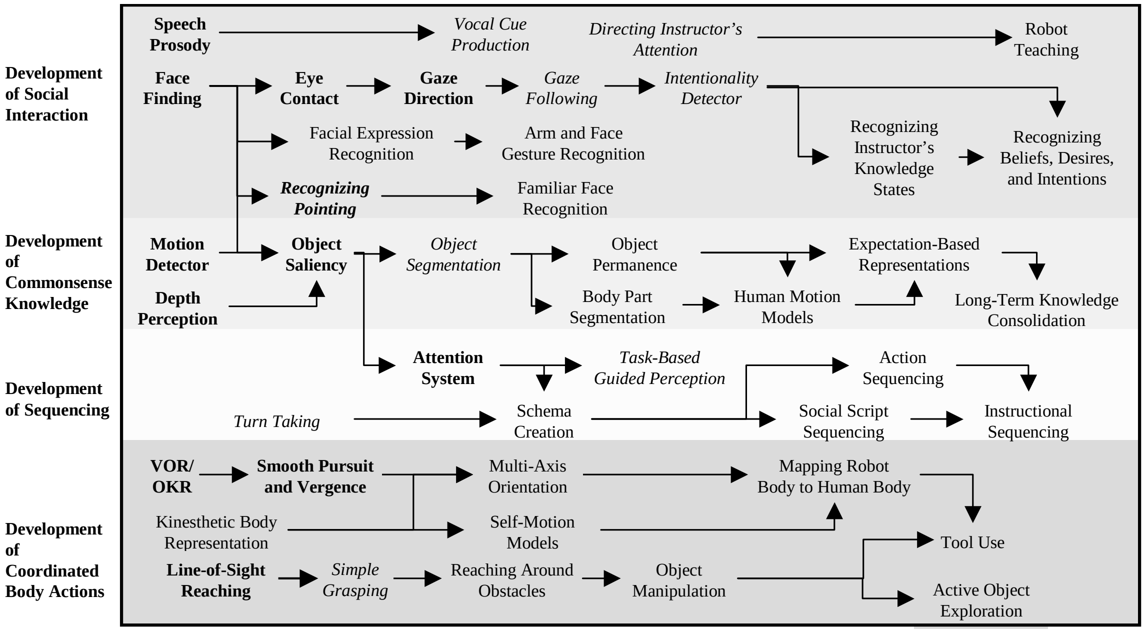

Development of Social Interaction
Development of Commonsense Knowledge
Development of Sequencing
Development of Coordinated Body Actions
Speech Prosody
Vocal Cue Production
Directing Instructor's Attention
Robot Teaching
Face Finding
Eye Contact
Gaze Direction
Gaze Following
Intentionality Detector
Facial Expression Recognition
Arm and Face Gesture Recognition
Recognizing Instructor's Knowledge States
Recognizing Beliefs, Desires, and Intentions
Recognizing Pointing
Familiar Face Recognition
Motion Detector
Object Saliency
Object Segmentation
Object Permanence
Expectation-Based Representations
Depth Perception
Body Part Segmentation
Human Motion Models
Long-Term Knowledge Consolidation
Attention System
Task-Based Guided Perception
Action Sequencing
Turn Taking
Schema Creation
Social Script Sequencing
Instructional Sequencing
VOR/ OKR
Smooth Pursuit and Vergence
Multi-Axis Orientation
Mapping Robot Body to Human Body
Kinesthetic Body Representation
Self-Motion Models
Tool Use
Line-of-Sight Reaching
Simple Grasping
Reaching Around Obstacles
Object Manipulation
Active Object Exploration

Figure 9-1: A sample of behavioral skills that are relevant to the study of imitation in robotics. This incomplete set represents behaviors that have been implemented (bold text), that have been partially implemented or implemented in a basic form (italic text), or are currently under investigation (normal text) by the Humanoid Robotics Group at MIT.

work in social learning for autonomous robots addressed learning inter-personal communication protocols between similar robots (Steels, 1996), and between robots with similar morphology but which differ in scale (Billard & Dautenhahn, 1998). Robotics research has also focused on how sequences of known behaviors can be chained together based on input from a model. Matarić et al. (1998) used a simulated humanoid to learn a sequence of gestures from a set of joint angles recorded from a human performing those same gestures, and Gaussier et al. (1998) used a neural network architecture to allow a robot to sequence motor primitives in order to follow the trajectory of a teacher robot. One research program has addressed how perceptual states can be categorized by matching against models of known behaviors; Demiris & Hayes (1999) implemented an architecture for the imitation of movement on a simulated humanoid by predictively matching observed sequences to known behaviors. Finally, a variety of research programs have aimed at training robots to perform single tasks by observing a human demonstrator. Schaal (1997) used a robot arm to learn a pendulum balancing task from constrained visual feedback, and Kuniyoshi et al. (1994) discussed a methodology for allowing a robot in a highly constrained environment to replicate a block stacking task performed by a human in a different part of the workspace.

Traditionally in robot social learning, the model is indifferent to the attempts of the observer to imitate it. In general, learning in adversarial or indifferent conditions is a very difficult problem that requires the observer to decide who to imitate, what to imitate, how to imitate, and when imitation is successful. To make the problem tractable in an indifferent environment, researchers have vastly simplified one or more aspects of the environment and the behaviors of the observer and the model. Many have simplified the problem by using only simple perceptions which are matched to relevant aspects of the task, such as Kuniyoshi et al. (1994), who use white objects on a black background without any distractors or Matarić et al. (1998), who place reflective markers on the human's joints and use multiple calibrated infrared cameras. Others have assumed the presence of a single model which is always detectable in the scene and which is always performing the task that the observer is programmed to learn, such as Gaussier et al. (1998), and Schaal (1997). Many have simplified the problem of action selection by having limited observable behaviors and limited responses (such as Steels, 1996; Demiris & Hayes, 1999), by assuming that it is always an appropriate time and place to imitate (such as Dautenhahn, 1995), and by fixing the mapping between observed behaviors and response actions (such as Billard & Dautenhahn, 1998). Few have addressed the issue of evaluating the success of an imitative response; most systems use a single, fixed success criteria which can only be used to learn a strictly specified task with no hope for error recovery (although see Nehaniv & Dautenhahn, 1998, for one treatment of evaluation and body mapping).

Our approach is to constrain the learning scenario in a different manner - we assume that the model is motivated to help the observer learn the task. A good teacher is very perceptive to the limitations of the learner and sets the complexity of the instruction and task accordingly. As the learner's performance improves, the instructor incrementally increases the complexity of the task. In this way, the learner is always competent but slightly challenged - a condition amenable for successful learning. This assumption allows us to build useful implementations on our robots, but limits the applicability of these results to less constrained learning environments (such as having an indifferent model). However, we believe that the problems that must be addressed in building systems with the assumption of an active instructor are also applicable to robotics programs that use other assumptions and to investigations of social learning in natural systems.

Evaluating complex robotic systems presents another level of challenges. Most individual components can be evaluated as stand-alone modules using traditional engineering performance measures, such as comparisons against standardized data sets or considerations of optimization and efficiency. Evaluating the behavior of an integrated system using standard techniques from ethology and behavioral psychology is difficult for many reasons. First, before the complete behavior can be evaluated, all of the required system components must be implemented and integrated together. Second, the particular assumptions used in constructing the system may limit the types of interactions that the robot can be evaluated under. For example, limits to perception may restrict the robot to only certain limited classes of stimuli, or to stimuli that are marked in certain ways. Similarly, simplified sets of motor responses can limit the types of behavior that we can expect to observe. Third, long-term studies of behavior are difficult because the hardware systems are fragile and constantly changing. Simply maintaining a robot at a given level of functionality requires full-time support, and few robotic systems are designed to operate for extended periods of time without human intervention. Furthermore, because of the expenses of building a robot, each research robot is often supporting a variety of research studies, many of which are constantly altering the hardware platform. Fourth, comparing results between robots is difficult because of differences in the underlying assumptions and in the hardware platforms. Despite these difficulties, we believe that the application of behavioral measurement techniques will be a critical step in the development of future robots. It is a goal of our research to achieve a level of functionality with our robots that would permit such an evaluation.

9.3 Challenges in Building Robots that Imitate People

The ability to imitate relies upon many perceptual, cognitive, and motor capabilities. Many of these requirements are precursor skills which are necessary before attempting any task of this complexity, but which are not directly related to the act of imitation. For example, the robot will require systems for basic visual-motor behaviors (such as smooth pursuit tracking and vergence), perceptual abilities for detecting motion, color, and scene segmentation, postural control, manipulative abilities such as reaching for a visual target or controlled-force grasping, social skills such as turn taking and recognition of emotional states, as well as an intuitive physics (including object permanence, support relations, and the ability to predict outcomes before attempting an action).

Even if we were to construct a system which had all of the requisite precursor skills, the act of imitation also presents its own unique set of research questions. Each of these questions is a complex problem which the robotics community has only begun to address. In this section, we focus on four of these questions:

- How does the robot know when to imitate?
- How does the robot know what to imitate?
- How does the robot map observed actions into behavioral responses?
- How does the robot evaluate its actions, correct errors, and recognize when it has achieved its goal?

To investigate these questions, consider the following example: A robot is observing a model opening a glass jar. The model approaches the robot and places the jar on a table near the robot. The model rubs his hands together and then sets himself to removing the lid from the jar. He grasps the glass jar in one hand and the lid in the other and begins to unscrew the lid. While he is opening the jar, he pauses to wipe his brow, and glances at the robot to see what it is doing. He then resumes opening the jar. The robot then attempts to imitate the action. We will use this example throughout this section to demonstrate some of the unique challenges to mimicry and imitation.

9.3.1 How Does the Robot Know When to Imitate?

A socially intelligent robot should be able to use imitation for the variety of purposes that humans do. Human children use imitation not only to acquire new skills, but also to acquire new goals from their parents. By inferring the intention behind the observed actions, children can gain an understanding of the goals of an individual. Children also use imitation to acquire knowledge about socializing, including the social conventions of their culture and the acceptable dynamics necessary for social communication. Imitation can be a mechanism for developing social attachments through imitative play and for gaining an understanding of people. Just as infants learn about physical objects by acting on them, infants learn about people by interacting with them. As Meltzoff & Moore (1994) wrote, "Imitation is to understanding people as physical manipulation is to understanding things." Imitation can also be used to explore and expand the range of possible actions in the child's repertoire, learning new ways of manipulating objects or new motor patterns that the child might not otherwise discover. Finally, imitation can be a mechanism for establishing personal identity and discovering distinctions between self and other. Meltzoff & Moore (1994) have proposed that deferred imitation may serve to establish the identity of a previously encountered individual.

A social robot should selectively use imitation to achieve many of these goals. However, the robot must not merely be a "puppet on a string." The robot must decide whether or not it is appropriate to engage in imitative behavior based on the current social context, the availability of a good model, and the robot's internal goals and motivations. For example, the robot may need to choose between attending to a learning opportunity or fulfilling another goal, such as recharging its batteries. This decision will be based upon the social environment, how likely the robot is to have another opportunity to engage in that particular learning opportunity, the current level of necessity for charging the batteries, the quality of the instruction, and other competing motivations and goals. When faced with the example of opening a jar, the robot must identify that the person is attempting to demonstrate a skill that should be imitated and recognize when during that interaction it is appropriate to attempt a response. Furthermore, the robot should also recognize when imitation is a viable solution and act to bring about the social context in which it can learn by observation, perhaps by seeking out an instructor or motivating the instructor to perform a certain task.

9.3.2 How Does the Robot Know What to Imitate?

Faced with an incoming stream of sensory data, the robot must make a number of decisions to determine what actions in the world are appropriate to imitate. The robot must first determine which agents in the scene are good models (and be able to avoid bad models). The robot must not only be able to distinguish the class of stimuli (including humans and perhaps other robots) which might be a good model but also determine if the current actions of that agent are worthy of imitation. Not all humans at all times will be good models, and imitation may only be appropriate under certain circumstances.

Once a model has been selected, how does the robot determine which of the model's actions are relevant to the task, which may be part of the social/instructional process, and which are circumstantial? In the example above, the robot must segment the scene into salient objects (such as the instructor's hand, the lid, and the jar) and actions (the instructor's moving hand twisting the cap and the instructor's head turning toward the robot). The robot must determine which of these objects and events are necessary to the task at hand (such as the jar and the movement of the instructor's elbow), which events and actions are important to the instructional process but not to the task itself (such as the movement of the instructor's head), and which are inconsequential (such as the instructor wiping his brow). The robot must also determine to what extent each action must be imitated. For example, in removing the lid from a jar, the movement of the instructor's hand is a critical part of the task while the instructor's posture is not The robot must also recognize the important aspects of the objects being manipulated so that the learned action will be applied to only appropriate objects of the same class (Scassellati, 1999*b*).

9.3.3 How Does the Robot Map Observed Actions into Behavioral Responses?

Once the robot has identified salient aspects of the scene, how does it determine what actions it should take? When the robot observes a model opening a jar, how does the robot convert that perception into a sequence of motor actions that will bring its arm to achieve the same result? Mapping from one body to another involves not only determining which body parts have similar structure but also transforming the observed movements into motions that the robot is capable of performing. For example, if the instructor is unscrewing the lid of the jar, the robot must first identify that the motion of the arm and hand are relevant to the task and determine that its own hand and arm are capable of performing this action. The robot must then observe the movements of the instructor's hand and arm and map those movements into the motor coordinates of its own body.

9.3.4 How Does the Robot Evaluate its Actions, Correct Errors, and Recognize Success?

Once a robot can observe an action and attempt to imitate it, how can the robot determine whether or not it has been successful? In order to compare its actions with respect to those of the model, the robot must be able to identify the desired outcome and to judge how similar its own actions were to that outcome. If the robot is attempting to unscrew the lid of a jar, has the robot been successful if it merely mimics the model and rotates the lid but leaves the lid on the jar? Is the robot successful if it removes the lid by pulling instead of twisting? Is the robot successful if it smashes the jar in order to open it? In the absence of internal motivations that provide feedback on the success of the action, the evaluation will depend on an understanding of the goals and intentions of the model. Further, if the robot has been unsuccessful, how does it determine which parts of its performance were inadequate? The robot must be able to diagnose its own errors in order to incrementally improve performance.

9.4 An Approach to Building Imitative Systems

Our approach to building systems that address the problems of determining saliency and relevance, mapping observed actions into behavioral responses, and implementing incremental refinement focuses on three keystones. First, saliency results from a combination of inherent object qualities, contextual influences, and the model's attention. This provides the basis for building perceptual systems that can respond to complex social situations. Second, our robots use similar physical morphologies to simplify the task of body mapping and recognizing success. By building human-like robots, we can vastly simplify the problems of mapping perceived actions to behavioral responses while providing an interface that is intuitive and easy to correct. Third, our systems exploit the structure of social interactions. By recognizing the social context and the stereotypical social actions made by the model, our robots can recognize saliency. By engaging in those same types of stereotypical social actions, the dynamics between the robot and the model provide a simplified means for recognizing success and diagnosing failures.

9.4.1 Multiple Sources of Saliency

Knowing what to imitate is fundamentally a problem of determining saliency. Objects can gain saliency (that is, become the target of attention) through a variety of means, including inherent object qualities, contextual influences and the model's attention. At times, objects are salient to people and animals because of their inherent properties; objects that move quickly, objects that have bright colors, and objects that are shaped like faces are all likely to attract attention. (We call these properties inherent rather than intrinsic because they are perceptual properties, and thus are observer-dependent and not strictly a quality of an external object.) Objects can also become salient through contextual effects. The current motivational state, emotional state, and knowledge of the observer can impact saliency. For example, when the observer is hungry, images of food will have higher saliency than they otherwise would. Objects can also become salient if they are the focus of the model's attention. For example, if the model is staring intently at a glass jar, the jar may become a salient part of the scene even if it is otherwise uninteresting. Fundamental social cues (such as gaze direction) can also be used by the observer to determine the important features of a task. People naturally attend to the key aspects of a task while performing that task. For example, when opening the jar, the model will naturally look at the lid as he grasps it and at his own hand while twisting off the lid. By directing its own attention to the object of the model's attention, the observer will automatically attend to the critical aspects of the task. In the case of social instruction, the observer's gaze direction can also serve as an important feedback signal for the instructor. For example, if the observer is not attending to the jar, then the instructor can actively direct the observer's attention by increasing the jar's saliency, perhaps by pointing to it or tapping on it.

9.4.2 Similar Physical Morphologies

Three of the problems outlined above can be simplified by assuming a similar physical morphology between the model and the observer. If the observer and model have a similar shape, the perceptual task of determining saliency can be constrained by the possible actions of the observer. If the observer witnesses an ambiguous motion of the model's arm, the observer can postulate that the perception must have been one of the actions which it could possibly perform in that situation and eliminate any other possible perceptual interpretations.

The mapping problem can also be simplified by having similar physical morphologies. If the observer can identify that it is the model's arm that is moving, it need not initially try to match that motion with an action that it is capable of performing only with its mouth or legs. Additionally, the position of the model's arm serves as a guideline for an initial configuration for the observer's arm. A different morphology would imply the need to solve an inverse kinematics problem in order to arrive at a starting position or the more complicated problem of mapping unlike body parts between model and observer (for example, see Herman, 2001, for imitation between dolphins and humans). In general this transformation has many solutions, and it is difficult to add other constraints which may be important (e.g., reducing loading or avoiding obstacles). By constraining the space of possible mappings, the computational complexity of the task is reduced.

Similar physical morphology also allows for a more accurate evaluation. If the observer's morphology is similar to the model's, then the observer is likely to have similar failure modes. This potentially allows the observer to characterize its own failures by observing the failures of the model. If the observer watches the model having difficulty opening the jar when his elbows are close together, the observer may be able to extrapolate that it too will fail without sufficient leverage. In situations where the model is taking an active role in instructing the observer, a similar morphology also allows the model to more easily identify and correct errors from the observer. If the observer's arms are too close together when attempting to open the jar, the model's knowledge about his own body will assist him in evaluating the failure mode and in providing an appropriate solution.

9.4.3 Exploit the Structure of Social Interactions

Social interactions have structure that can be exploited to simplify the problems of imitation. By recognizing the appropriate social context, the observer can limit the number of possible perceptual states and determine whether the attention state of the model is an appropriate saliency signal. When the model is performing a manipulative task, the focus of attention is often very relevant. However, when engaged in some social contexts, the focus of attention is not necessarily important. For example, it is customary in many cultures to avert eye contact while taking one's turn in a conversation and to establish eye contact when ending a turn. Exploiting these rules of social conduct can help the observer to recognize the possible value of the attention state of the model (thus simplifying the saliency problem).

The structure of social interactions can also be used to provide feedback in order to recognize success and correct failures. In the case of social instruction, the difficulty of obtaining success criteria can be simplified by exploiting the natural structure of social interactions. As the observer acts, the facial expressions (smiles or frowns), vocalizations, gestures (nodding or shaking of the head), and other actions of the model all provide feedback that will allow the observer to determine whether or not it has achieved the desired goal. The structure of instructional situations is iterative; the instructor demonstrates, the student performs, and then the instructor demonstrates again, often exaggerating or focusing on aspects of the task that were not performed successfully. The instructor continually modifies the way he performs the task, perhaps exaggerating those aspects that the student performed inadequately, in an effort to refine the student's subsequent performance. By repeatedly responding to the same social cues that initially allowed the observer to understand and identify which salient aspects of the scene to imitate, the observer can incrementally refine its approximation of the actions of the instructor.

Monitoring the structure of the social interaction can assist the instructor in maintaining an appropriate environment for learning. Expressive cues such as facial expressions or vocalizations can regulate the rate and quality of instruction. The instructor modifies both the speed and the content of the demonstration based on feedback from the student. By appearing confused, the student causes the instructor to slow down and simplify the demonstration.

Recognizing the appropriate social context can be an important cue in knowing when imitation is an appropriate solution to a problem. Internal motivations will serve as a primary mechanism for determining when to search for an appropriate model and when an attempt to perform an imitative act is appropriate. However, opportunistic use of good models in the environment can also be important in learning new skills. By recognizing which social contexts are likely to produce a good model behavior, the robot can exploit learning opportunities when they arise.

9.5 A Mechanism for Mimicry

Using the perceptual and motor systems developed in previous chapters, a basic set of mimicry behaviors can be constructed. As an initial proof of concept, any interesting trajectory will be mapped directly to a trajectory of the arm as described in section 5.3.2 (this mapping was a joint research project with Bryan Adams). The attention system will serve as the primary arbiter of saliency, integrating properties from in-

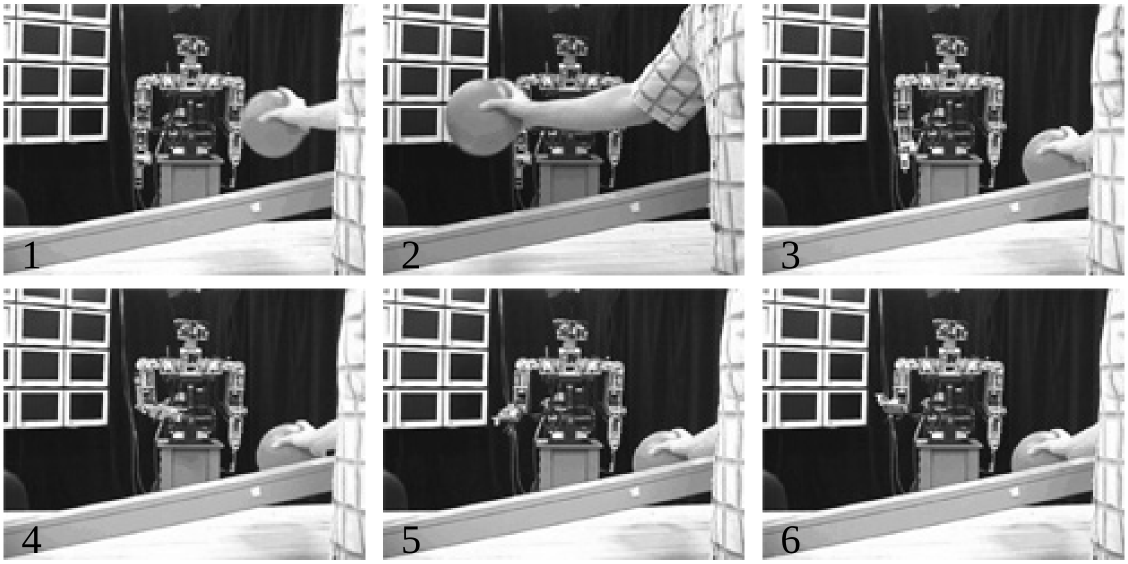


Figure 9-2: Still images from a video recording of a simple demonstration of mimicry. The frames shown were spaced apart by approximately one second in the original video. Images 1–3 show a person moving a bright red ball in a figure eight pattern. The robot observes this action and then responds by making a similar movement with its right arm.

herent object properties, the attentional state of the instructor, and the motivational constraints. Trajectories will be formed by the multi-hypothesis tracking algorithm, and the animacy judgment from ToBY (chapter 7) will serve to select appropriate trajectories. This process will be described in further detail by considering a set of examples. For each of the examples, a person interacted with Cog from a distance of between two and eight feet. There were an additional two to six people in the room during each of these interactions, but none were actively trying to engage the robot. Interactions were video recorded, and still images were extracted from the video sequence. The sequence of still images were extracted at roughly regular intervals (every one or two seconds) but deviations of a few frames were allowed to provide clearer static images.

Figure 9-2 shows an example of the basic mimicry behavior. A brightly colored ball was moved in a figure eight pattern in front of Cog. The ball was detected by the attention system by virtue both of the motion and the high color saturation. The trajectory was linked by the multi-hypothesis tracking system and the agents of the theory of body module classified the movement as animate. The multi-hypothesis tracking algorithm was designed to produce trajectories of only 30-60 elements (1-2 seconds of data), but mimicry requires a representation of trajectories that is longer. This is easily accomplished, since the tracking algorithm places a unique identifier with each detected trajectory. A second processing stage simply collects the trajectory points for each unique trajectory, throwing away trajectories when they are no longer part of the visual scene. To maintain tractability, trajectory points that are older

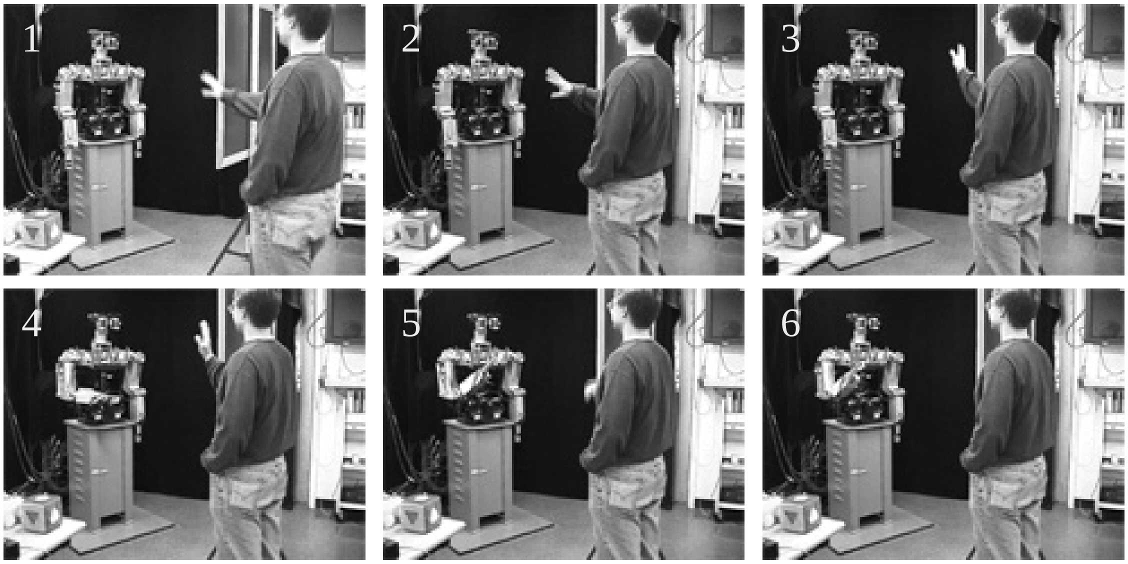


Figure 9-3: Images taken from a video clip showing Cog mimicking the author waving. Approximately one second of video separates each image. In this example, the range of motion and spatial position of the robot's response are determined by the scale and location of the person's face.

than 15 seconds are also discarded. For this most basic behavior, the points in the trajectory were mapped linearly into a two dimensional space with limits at ±1. Thus, a position (x, y) within an image of size $w \times h$, the remapped positions were at:

$$(x', y') = (\frac{2x}{w} - 1, \frac{2y}{h} - 1)$$

These positions were then used as the linear combination weights for the set of postural primitives defined in section 5.3.2. Because the sequence of points in the trajectory were regularly spaced in time, the command position for the robot's arm was updated by these remapped trajectory positions at 30 Hz. This allowed the robot to match not only the spatial form of the movement but also the temporal characteristics. The resulting behavior allowed the robot to replicate a movement that had a similar two-dimensional projection as the original movement of the model object. This full-field mapping is useful if there is no additional information about the scene, but additional information can provide a more interesting spatial reproduction.

A more complex representation of the spatial properties of the movement can be created if there is additional information about the creation of the model trajectory. For example, if the movement results from a person's hand, the size and position of the hand with respect to the body are relevant for determining the extent of the movement. Figure 9-3 shows an example of this spatial mapping. In this example, a person waving at the robot from a distance of approximately six feet produced a small retinal projection (the trajectory ranged over 30 columns and 20 rows in the 128×128

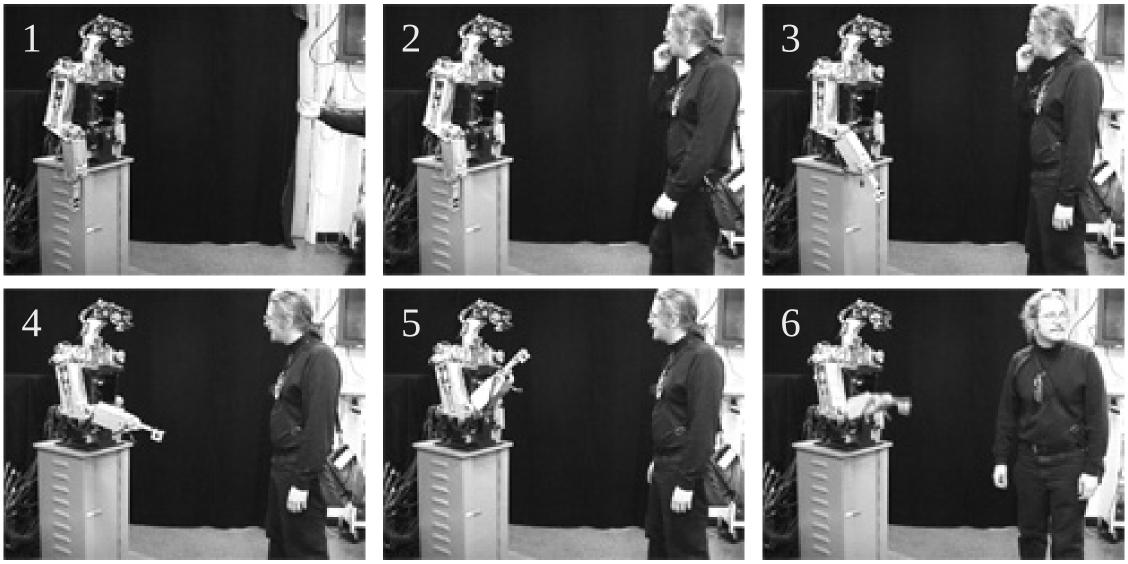


Figure 9-4: Cog mimicking a person making the American Sign Language sign for food. The person began making the gesture before entering the field of view of the video camera, but the length of time of each gesture was still similar.

image) near the center of the image. Using the full-field mapping, this would result in a relatively small movement of the robot's arm near the primary primitive posture (the origin of the postural primitive space). This would result in an arm movement that varied slightly around this primary posture, which was originally set with the elbow near the robot's side and the forearm straight out in front of the body, parallel to the floor. With the aid of a complete perceptual model of the body structure of the individual being observed, a more accurate reconstruction could be achieved; if the robot could compute the relative positions of the joints of the human's arm, replicating that structure would be simple. While there are computational systems that attempt to extract body posture from video sequences (for example Rosales & Sclaroff, 1999; Darrell et al., 1998*b*), these techniques are often computationally expensive or require a simplified visual scene or a structured background. As an alternative, we have constructed a mapping that uses the scale of detected faces in the image as an indication of the range of motion that should be performed. A detected face centered at the image coordinates (x_f, y_f) with a size of $w_f \times h_f$ is assumed to be structured around a body centered at (x_b, y_b) with a size of $w_b \times h_b$. To map upper-body movements to the range of postural primitives, the following values were chosen:

$$w_b = 6 \times w_f \tag{9.1}$$

$$h_b = 4 \times h_f \tag{9.2}$$

$$x_b = x_f \tag{9.3}$$

$$y_b = y_f + 1.5 \times h_f \tag{9.4}$$

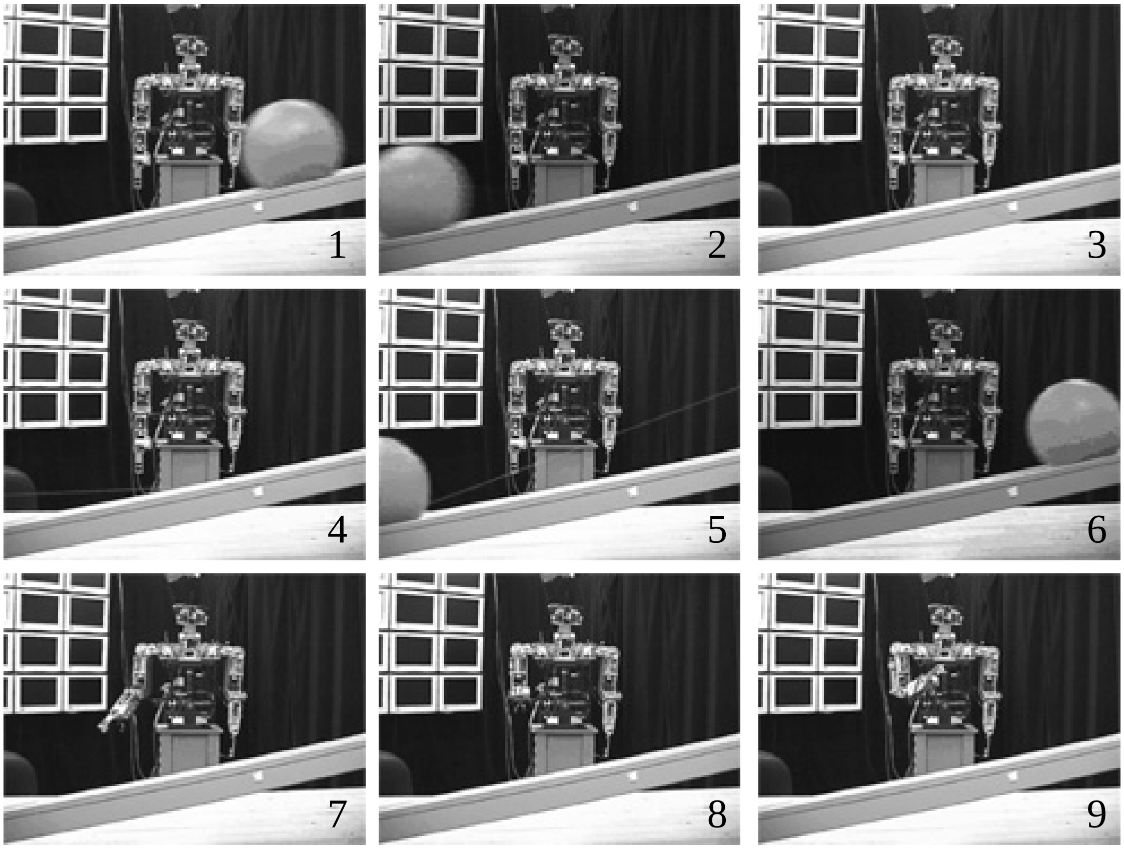


Figure 9-5: The mimicry system can also be made to depend on ToBY by allowing imitation only of animate trajectories. In this example, the robot observes a ball rolling down a slope (images 1-2), which ToBY classifies as inanimate. The robot does not respond to this inanimate trajectory (images 3-4). When the same ball is pulled up the inclined plane by a piece of fishing wire (images 5-6), the robot responds by mimicking the movement of the ball (images 7-9). Each frame shown here was separated by approximately two seconds of video.

If more than 75% of a selected trajectory exists within the bounding box of a body range of a detected face, then this body range is used as the full range for transforming the trajectory into postural primitive coordinates. Otherwise, the full range of the image is used (as was the case in the previous example). The result of this mapping, as seen in figure 9-3, allows the robot to more closely approximate human movements. Figure 9-4 shows a similar mapping with another individual. The robot maps the movement of the hand to a range relative to its own body that matches the range of motion of the person's hand relative to his body.

In the past two examples, it was assumed that the appropriate trajectory had been selected for imitation. In a social setting, the problem of choosing the appropriate trajectory is complex. With many people in the robot's field of view, there are always a variety of visual targets. Simple mechanisms can quickly rule out some background trajectories. Trajectories that are too short (have fewer than thirty points) or that have a cumulative displacement that is below a threshold (generally, 25 pixels) are quickly rejected. However, additional criteria such as selecting the trajectory with the greatest spatial displacement (or maximum velocity) proved to be too simplistic. The classification performed by the theory of body module can serve as an excellent discrimination function for selecting the most interesting trajectory for mimicry. Only trajectories that are classified as animate are considered as candidates for mimicry. Figure 9-5 shows an example of this discrimination. For this demonstration, the trajectory system and the imitation system were applied to the foveal camera image, rather than the peripheral camera image. This allowed the same object and spatial conditions to be used to describe both an animate and inanimate trajectory without the construction of a large testing apparatus. Because the robot has no concept of object permanency, when an object leaves the field of view and re-enters, it is considered to be a completely different trajectory. The video recorder was placed such that the field of view of the robot's foveal camera matched the recorder's field of view; when an object disappears from the video, it also disappeared from the robot's sight. As a simple demonstration, a brightly colored ball was allowed to roll down a ramp (images 1-2 of figure 9-5). The robot classified this movement as inanimate and did not respond (images 3-4). When the ball was pulled up the same incline using an attached string (images 5-6), this trajectory was classified as animate and the robot responded by mimicking the movement (images 7-9). While this demonstration was certainly artificial, the ToBY classifications were critical in removing background trajectories in complex social environments.

To evaluate the mimicry system in natural social environments, groups of naive subjects were recruited to interact with the robot.[3] Twelve subjects, ranging in age from 7 years to 39 years, were brought into the lab either in pairs (two pairs) or individually (eight subjects) and asked to interact with the robot. (Most subjects were quite happy to just go and "play" with the robots without any additional instructions.) The subjects were given no explanation of the range of behaviors that the robots could perform, nor were they told the robot's perceptual abilities. Subjects had access to a variety of brightly colored children's toys and a variety of other objects that are common in our laboratory. Because this was also a pilot study, subjects were observed by four to six researchers while they were interacting with the robot. While the observers never directly engaged the robot, they did act as distractors by providing a wide assortment of background movement (both animate and inanimate). While subjects were in the laboratory, Cog performed three different behaviors. First, the robot's head and neck would saccade and orient to salient targets. Second, when the most salient object in the scene had a very high color saturation, the robot would point/reach toward that object. Third, the robot would mimic movements that were animate. The head and eye movement was driven directly be the attention system,

[3] A more complete study of the forms of social interactions that occur between naive users and the robots Cog and Kismet is currently being performed by Jen Audley, Cynthia Breazeal, Brian Scassellati, and Sherry Turkle. Some of the subjects reported here were also used as pilot subjects for these more extensive studies.

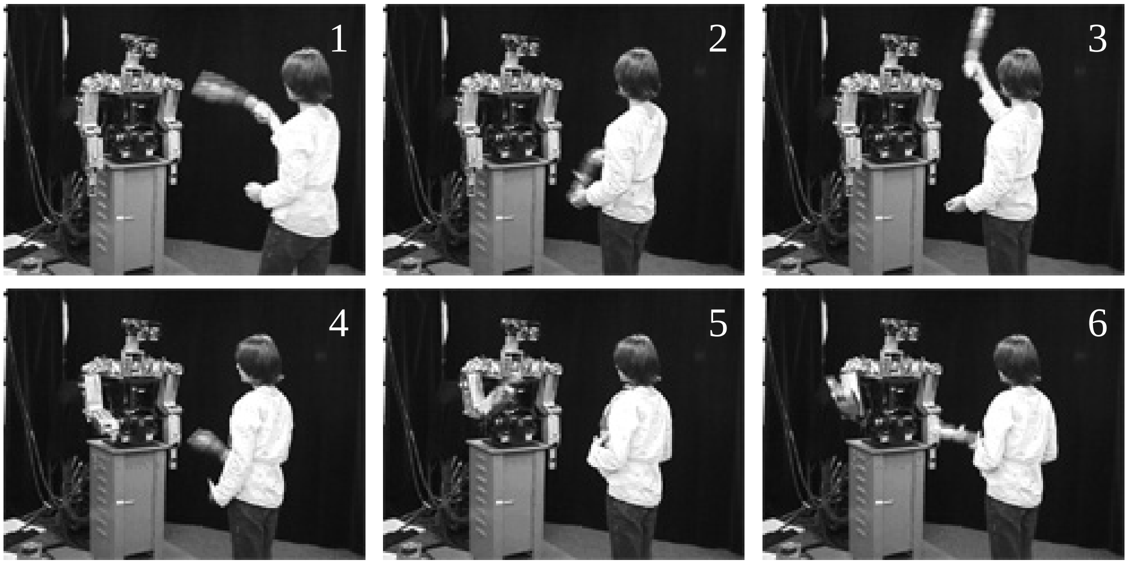


Figure 9-6: Additional testing was performed with subjects who knew nothing about the capabilities of the robot. A naive subject swings a toy inchworm clockwise in front of the robot in an attempt to attract the robot's attention (images 1-3). When the robot responds by making a similar circular motion with its arm (images 4-6), the subject said to the robot "wow, that's pretty cool...now try this."

and did not interact directly with the arm movement. A simple arbiter process mediated between the two arm behavior movements on a first-come first-served basis; while one behavior was active, the other was actively suppressed.

The effectiveness of the mimicry system was evaluated both while the naive subjects were unaware of the behaviors that the robot could perform and under a directed interaction with the robot. Of the twelve subjects, eleven subjects gave a verbal report within the first five minutes of interacting with the robot that indicated that they were aware that the robot was mimicking their movements. For example, one thirteen year-old subject noted "hey, it's doing what I'm doing." The twelfth subject also clearly noticed the correlation and began engaging the robot in an imitation game but gave no verbal report of what was happening. While subjects were uniformly interested in getting the robot's attention and in engaging the robot socially, the discovery of the imitative behavior was always accompanied by a heightened level of excitement. Many subjects also began to use this mimicry behavior as a way of discovering the robot's capabilities by trying to find the range of its movement or the range of stimuli that the robot found salient. Figure 9-6 shows one of these subjects swinging a plush toy inchworm in front of Cog. She swings the toy in a large clockwise circle (images 1-3), and the robot responds by swinging its arm in a similar circle (images 4-6). The subject then said to the robot, "wow, that's pretty cool...now try this." That subjects could spontaneously recognize the arm movements as mimicry even in the presence of other similar arm movements gives a good indication that they were socially engaged by this process. Once subjects had become aware that the robot was mimicking their movements, we further asked them to deliberately get the robot to perform certain actions such as waving goodbye, reaching out to the side, making a circular movement, or a Zorro-like swash in front of it. All subjects in this case were able to attract the robot's attention and to get the robot to mimic their movement. At times, this often involved the performance of an action more than once, and on some occasions subjects actively attempted to improve the performance of the robot by providing further verbal cues, by exaggerating their movements, or by presenting a more salient or easily processed visual cue. The success of these subjects at performing these tasks (and at manipulating the robot to perform tasks) demonstrates that the robot is perceiving at least some of the social cues that people find natural to use in these situations.

There are many obvious extensions of this mimicry system: recognition of vocal cues as a feedback mechanism, a perceptual system that does analysis of articulated movement, three-dimensional representations of object trajectories, and many others. The following two chapters will each introduce an additional perceptual criteria that enables a more complex and responsive mimicry system within the context of the embodied theory of mind model. One further extension that is currently under development in our laboratory also deserves mention here. Edsinger (2001) has been using the perceptual systems described in this work (the attention system, trajectory formation, and the ToBY trajectory analysis) to demonstrate a mimicry system that includes a more complex sensorimotor mapping. Rather than mapping visual trajectories to a fixed coordinate frame of postural primitives, Edsinger has defined a set of behavioral actions and uses a spline-based representation to map between observed trajectories and their behavioral counterparts. In many ways, this may be one step closer to the mechanisms for behavioral responses in humans and other animals. By mapping observed states onto a finite set of well-known behaviors, Edsinger (2001) can accomplish more natural, optimized movements.

Chapter 10

Shared Attention Mechanisms

> *Thus, it is possible that human embodiment supports joint attention, joint attention supports gesture, gesture supports the representational properties of language, and symbolic language supports the capacity for mentalising. Such a developmental sequence need not carry any implication that since the mind is supposedly unobservable it therefore has to be theorized.* – Butterworth (2000)

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 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 caregiver in order to learn about their environment, in part by allowing the infant to manipulate the behavior of the caregiver and in part by providing a basis for more complex forms of social communication such as language and gestures (Lund & Duchan, 1983; Baldwin, 1991).

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 king-

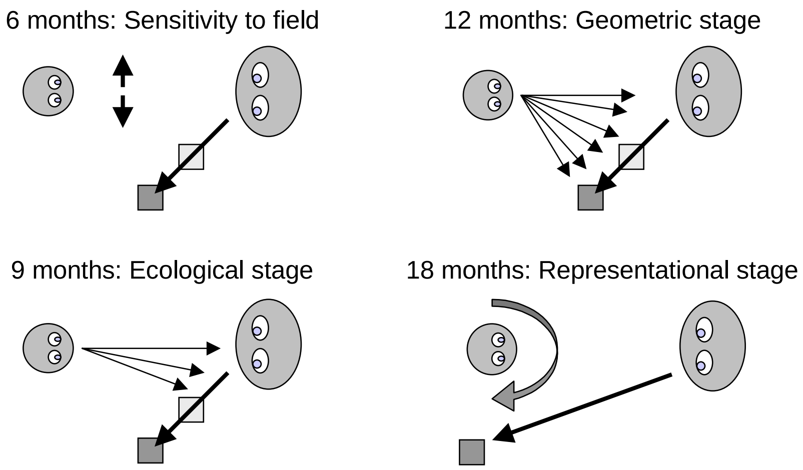

6 months: Sensitivity to field
12 months: Geometric stage
9 months: Ecological stage
18 months: Representational stage

Figure 10-1: Stages in the development of joint reference proposed by Butterworth (1991). Children initially are sensitive only to the left/right direction of gaze of the parent. By nine months, the child is capable of projecting along a rough directional vector from the adult's gaze direction, but tend to stop at the first inherently salient object along that scan path. Around 12 months, the child correctly interprets the direction of gaze as a three dimensional reference, but will not turn to look at objects that are outside the field of view until 18 months.

dom 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).

Butterworth (1991) has conducted particularly detailed investigations of the development of joint reference and has proposed a four-stage model (see figure 10-1). Each of these stages can be demonstrated by observing the behavior of an infant when an adult who is making eye contact with the infant moves their gaze to another object. 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 caregiver looking to the left and the caregiver looking to the right. Over the next three months, the infant's accuracy increases, allowing a rough determination of 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 can successfully determine not only the angle of gaze but also the vergence of the eyes. However, even at this stage, infants will only exhibit gaze following if the distal object is within view while looking at the adult. Children of this age will not turn to look behind themselves, even if the angle of gaze from the caregiver 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.

Using the perceptual primitives for detecting head orientation developed in chapter 8, basic examples of joint reference can be constructed. Using Butterworth's first two stages as a guideline, section 10.1 will demonstrate how a relatively simple feedback system between the head pose orientation module and the attention system developed in chapter 6 can generate joint reference. This implementation will be supported with examples of biasing the mimicry behavior discussed in the previous chapter. Section 10.2 will discuss potential implications of this implementation on models (including Baron-Cohen, 1995) that maintain a strict modular representation of joint reference.

10.1 Implementing Joint Reference

A robot capable of engaging in joint reference behaviors with a human requires three types of capabilities: a physical structure that allows the human to attribute attentional states to the robot, a perceptual system that is capable of recognizing the social cues indicative of attention in the human, and the ability to link these perceptual states to behaviors that direct attention. The first two requirements are already present in the system design that has been presented thus far. Cog can produce the appropriate social cues of attention through eye and neck movements (orientation behaviors), through visual tracking, and through pointing gestures. These behaviors, combined with the robot's anthropomorphic appearance, are naturally interpreted by humans, even those who have no experience interacting with the robot. While Cog cannot recognize all of the complex perceptual signals involved in social interaction, recognition of head pose is a sufficient social cue to evoke joint reference behavior.

The one remaining requirement is to link this attentional state to behavior that directs the robot's attention. In the model of Baron-Cohen (1995), this purpose is served by SAM, the shared attention mechanism. In Baron-Cohen's terms, SAM is a "neurocognitive mechanism" rather than a module in sense of Fodor (1992). However, the treatment of SAM has always been as a distinct modular component – encapsulated knowledge that can be selectively present or absent. In the implementation discussed here, joint reference is not explicitly represented as a modular component. Rather, it is a property of a feedback mechanism between the head pose detection system and the attention system. This feedback loop, combined with the existing behavioral systems, produces the same joint reference behaviors as would be generated by SAM.

To complete the feedback between the perceptual processes that detect salient social cues and the behavioral systems that produce attentive behavior, a simple transformation must be employed. By modifying the fidelity of this transformation, the first three of Butterworth's stages of joint reference development can be achieved, although due to perceptual limitations only the first two will be demonstrated here. The output of the head pose detection system is a data structure that includes the location of the face, the scale of the face, and the orientation of the head in terms of yaw, pitch, and roll. The inputs to the attention system are all structured in terms of a retinotopic map. To achieve Butterworth's first stage (sensitivity to the field

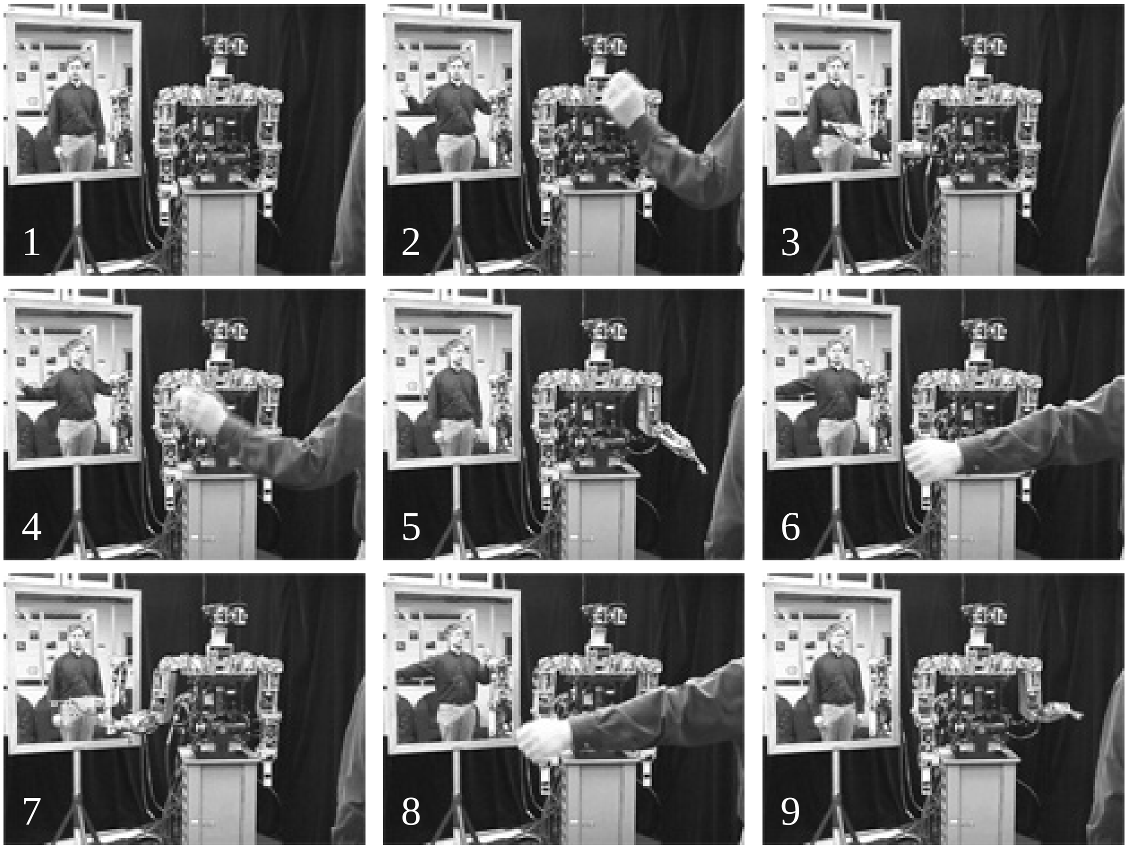


Figure 10-2: Nine frames from a video sequence showing the application of joint reference for selection of trajectories for mimicry. In this video, a large mirror was positioned behind the robot, outside its field of view, to permit the video camera to record both the actions of the human and the robot. When the human looks to the left and makes two arm movements (images 1-2), the robot responds by selecting an arm movement that matches the head orientation (image 3). Similarly, when the human looks to the right (image 4), the trajectory to the right becomes more salient, and the robot acts upon it by moving its left arm (image 5). Images 6-9 show the same effect for two arm movements that differ from each other. Approximately two seconds of video separated each of these images.

of view), the transformation marks all pixels to the appropriate side of the head as salient and all other pixels as uninteresting. For example, for a face located at row 20 and column 30 that is looking to the right (toward the origin), all pixels in columns 0-29 would received a value of 255, while all other pixels would receive a value of 0. (Recall that high values in the activation maps indicate salient regions.) To achieve the ecological stage of gaze following, a different mapping function is employed. The area of attention is modeled as a cone of attention that originates at the center of the face location and extends along an angle that matches the projection of the head orientation. To match the behavior of the ecological stage, the intensity of the cone is at a maximum (a pixel value of 255) at its origin and degrades by 10% every fifteen pixels of distance from the origin. This gives both a directional differential and a distance differential which biases the robot to attend to the first salient object along that scan path. In practice, a cone with an extent of 15 degrees to either side of the orientation angle was found to be effective.

The geometric stage can also be achieved with this method by using the same cone of attention but rather than degrading the intensity of the cone based on the distance from the origin, the intensity is degraded based on the distance from the perceived vergence target. In this way, targets at a specific distance from the observed person are enhanced. This capability has not been demonstrated on Cog because the perceptual processing is not sophisticated enough to recognize vergence angles or more detailed 3-D representations of pointing gestures. Similarly, a true representational stage of joint reference relies on the presence of other cognitive abilities for representing objects and for building representations of space that are not currently within the field of view, both of which are not currently implemented. A true representational stage would likely also directly influence search behaviors at a higher level than these pre-attentive processes.

The addition of a joint reference input to the attention system is not a capability originally envisioned by Wolfe (1994). While there is little evidence that these joint reference behaviors are at the same perceptual level as the other pre-attentive filters in human visual behavior, this implementation choice is a simple method to allow all of the robust behaviors that had previously been designed to act on the output of attentional processes to be driven by joint reference without the introduction of any additional mechanisms. The relative influence of joint reference can easily be modified simply by changing the weighting that is applied to that input channel in the attentional process.

In addition to driving attentional responses such as orientation and pointing behaviors, the effect of joint reference can also be applied to select appropriate trajectories to mimic. People tend to pay close attention to their movements and manipulations of objects. When attempting to instruct another individual, this tendency is even more pronounced. In this way, attention acts as a natural saliency cue by pointing out the important aspects of the social scene. On Cog, the integration of the joint reference cues into the attention system allows for the selection of salient trajectories based on joint reference to be implemented without any further software. Figure 10-2 shows an example of the influence of head orientation on mimicry. To allow both the robot's behavior and the human's behavior to be captured using only a single video camera, a large mirror was placed behind the robot. The robot could neither see nor reach the mirror. The human instructor then made either identical movements with both arms (images 1-5) or different movements with both arms (images 6-9) while looking and orienting either toward his own left (images 1-3 and 6-7) or right (images 4-5 and 8-9). To allow an easily observable behavioral difference, the robot was programmed to respond either with its left or right arm, depending on whether the robot selected a trajectory that was to the right or the left of a detected face. (Note that to act like a mirror image reflection, when the human acts with his left hand, the robot must respond with its right hand.) As figure 10-2 demonstrates, the response of the robot to joint reference cues can easily be reflected in the mimicry behavior.

10.2 Shared Attention without a Modular Structure

One of the primary differences between the embodied theory of mind presented here and the original work of Baron-Cohen (1995) is that the role of joint reference is not encapsulated within a single modular structure. The model presented here should not be taken as any sort of proof that the human system operates in the same way. It does however provide an existence proof that joint reference behavior can be produced without the need for a complex, encapsulated module. The embodied model provides a useful interface to behavior selection and can account for many of the basic properties observed in the development of joint reference skills in infants. This perspective is not unheard of within the developmental science community. In fact, shortly before his death, Butterworth (2000) had begun to articulate a position that joint attention is based on the properties of system embodiment. Butterworth noted that aspects of the design of the human body allowed the social cues that indicate attentional states to be more easily perceived. For example, the white color of the human sclera makes determining gaze direction easier.[1] He concluded that "it is possible that human embodiment supports joint attention, joint attention supports gesture, gesture supports the representational properties of language, and symbolic language supports the capacity for mentalising. Such a developmental sequence need not carry any implication that since the mind is supposedly unobservable it therefore has to be theorized." We agree with Butterworth that joint reference is supported by the basic facts of embodiment and that it can be grounded in perceptual states without resorting to wholly cognitive explanations of behaviors.

Although the system developed here uses only a single social cue for joint reference (head orientation), this architecture can easily be extended to deal with more complex perceptions and more complex behavioral responses. Gaze direction could be integrated into the attentional system using the same types of functional feedback

[1] It is also interesting to note that no other primate has this dramatic difference in coloring between the sclera and the iris and pupil.

connections as are currently used for head orientation. If a perceptual system were able to recognize body postures and could determine pointing gestures, this information could also be integrated using a similar set of rules for determining a cone of attention. More complex behavioral selection procedures could also allow the attentional states to directly trigger specific behavioral responses. For example, were the robot to observe a person pointing to an apple, the robot might also point to the apple (another case of mimicry) or might attempt to re-direct the attentional state of the person to an object that the robot was more interested in acquiring (thus interpreting the pointing gesture as a question).

Chapter 11

Detecting Simple Forms of Intent

> *If you should encounter a mountain lion while hiking in the Sierra Nevada mountains of California, there are two things you must not do, according to the Mountain Lion Foundation: turn your back on the animal or run away. Either of these behaviors would trigger the lion's predatory chase behavior, transforming you from startled hiker into potential prey. It is possible to avoid becoming prey by denying the lion's perceptual system the cues that normally accompany being a mealtime animal. Knowing how other creatures categorize behavior based on motion cues could thus make the difference between life and death.* – Blythe et al. (1999, p. 257)

In addition to the interpretations of animacy provided by ToBY, simple motion cues can also provide information about intention. The classic studies of Heider & Simmel (1944) demonstrated that people naturally attributed a wide variety of intentional states to even simple geometric shapes that moved across a background. The attribution of goal and intent has a central role in both the theories of Leslie (1994, as part of ToMM-1) and Baron-Cohen (1995, as ID). Furthermore, the attribution of intent is a critical distinction between mimicry and true imitation, or, using the terminology of Byrne (1999), the difference between copying the form of an action and the structure of an action. The close ties between intent and imitation were studied by Meltzoff (1995), who demonstrated the ability of infants as young as 18 months of age to distinguish between the surface form of an action and the underlying goal of an action.

While many of these studies focus on the fact that people are willing to attribute intention to basic object motion, very few research programs have addressed questions about the nature of the basic properties that cause these attributions or even a classification of the types of attributions that are applied. One notable exception is the work of Blythe et al. (1999), who attempted to build a classification system for a set of basic intentional states for a pair of simulated insects. Their experiments focused on two simulated ants in an on-screen environment that had no other objects or obstacles, but that did have a reasonable model of physics (friction, inertia, etc.). Three networked computers were attached to this simulation. Subjects at the first two computer consoles could control the movement of the ants through simple keyboard and mouse-based interfaces. A subject at the third terminal had no control over the simulation, but could observe the movements of the two ants. In the first phase of their study, the two subjects controlling ants were given certain intentional roles that they were asked to play out using the ants, such as "pursuing," "evading," "courting," "fighting," or "playing." These two subjects were isolated in separate rooms and had no interaction with each other except through the simulation. The third subject was asked to characterize the intentional role of each ant as they observed the interaction. In the second phase of their study, they attempted to derive classifiers that could look at the raw trajectories of the ant movement and produce judgments of intentionality that matched the human judgments. Their results demonstrated that even relatively complex intentional attributions (such as "play") can be discriminated solely on the basis of spatio-temporal properties of rigid body movement and that this discrimination can be performed by an algorithm.

The studies of Blythe et al. (1999) were so successful in part because of the limited range of intentional choices presented to subjects and also because of the simplicity of the environment. In building a basic representation of intent for a humanoid robot, we have chosen to deal with a much more restricted set of intentional states that can be recognized in exchange for being able to process a more complex environment. This chapter will present a very basic system that attributes two states of intentional relation in similar ways to the intentionality detector in Baron-Cohen (1995). This implementation will recognize both attraction and repulsion relationships, which might also be characterized as intentions of approach/desire and escape/fear. This implementation will differ significantly from the work of Blythe et al. (1999) in two ways: attributions of intent will only be applied to agents that exhibit self-propelled motion and this perceived intentional state will be used directly to drive behavior.

11.1 Recognizing Attraction and Repulsion

The intentionality detector (ID) takes as input the labeled trajectories that are produced by the theory of body module. Unlike ToBY which operates on each trajectory independently, the intentionality detector is primarily concerned with the relationships between trajectories. The motion of one object (the actor) can be interpreted in an intentional framework only with respect to the position (or movement) of a second object (the target). Because these trajectories are based on the salient objects detected by the attention system, ID can only represent relationships between pairs of salient objects. While this does limit the full potential of the system, the effect of the restriction on the behavior of the system is minor. In one way, the limitation is useful in that it restricts the number of possible pair-wise trajectory comparisons that must be performed.

Attributions of intent are only permitted to trajectories that have been classified as animate by the theory of body module. In this way, many spurious comparisons between pairs of inanimate objects are never computed, and the attribution of intent is critically tied to the internal representations of agency. Only those stimuli that are

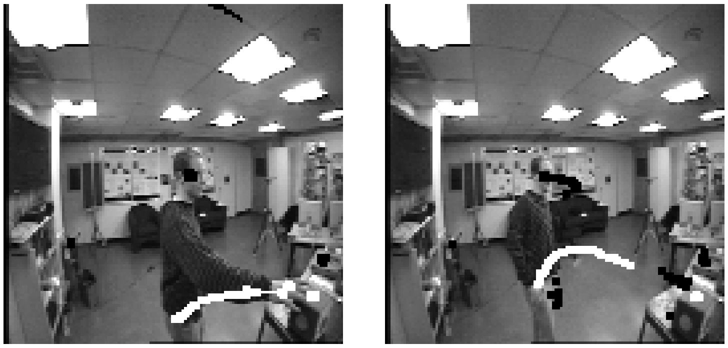


Figure 11-1: The intentionality detector seeks to label the intentional relationships for basic approach and withdrawal. The image at left shows a person reaching for a toy block with the trajectory positions drawn on the image. This trajectory was classified as an approach event. At right, an event that was classified as a withdrawal event when the person then quickly pulled his hand away.

classified as social agents are capable of acting as the primary actor in the intentional relations. Note that the target of the intentional relation may be either animate or inanimate; it is still possible for an animate agent (such as a person) to desire an inanimate object (such as an apple). It is also possible that the discrimination of animacy in the target stimulus might someday be a useful component in discriminating more complex forms of intent. For example, when I observe a man moving rapidly away from a woman, I might interpret that reaction differently than when I see a man moving rapidly away from a cardboard box.

The intentionality detection system implemented in this work classifies two types of intentional relationship: attraction and repulsion. While this set of relationships is certainly smaller than the set used in Blythe et al. (1999), these basic forms of attraction and repulsion may be the basis for more complex forms of intentional relation (Leslie & Keeble, 1987). Every animate trajectory is considered to be a possible actor. ID compares all possible pairings of possible actors with the complete set of trajectories (excluding the actor's own trajectory). To perform a comparison, the sets of points from each trajectory are aligned temporally by local shifting operations which match points in each trajectory that were acquired from the same original image frame. Each matched set of points is then compared using some simple spatial measurements such as distance, velocity, the difference in the angles of the velocity vectors (the relative heading), and the velocity angle of the actor with respect to the actual directional vector between the actor and the target (the approach vector). These statistics are identical to those used by Blythe et al. (1999) and are used directly to classify the intent as approach, avoidance, or neither. An approach relation occurs when:

- The difference in the relative headings of the actor and target, averaged over the number of points in the trajectories that could be aligned, is below a threshold of 20 degrees.

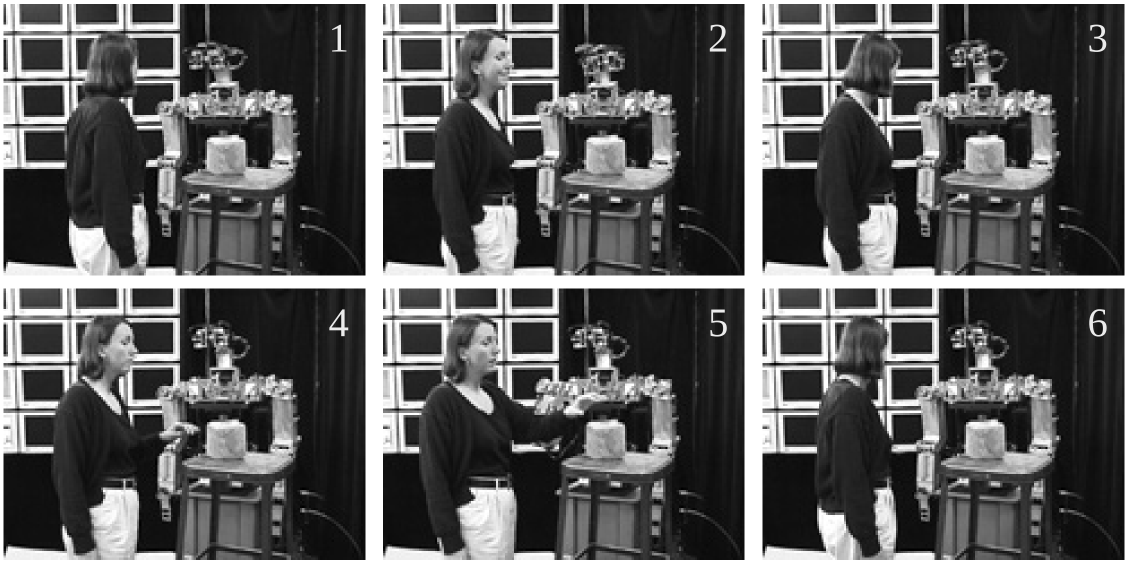


Figure 11-2: A demonstration of the use of intentional relationships to guide behavior. The subject was asked to perform a sequence of social tasks. First, the subject was asked to get the robot's attention (image 1). The subject was then asked to orient to a nearby toy (image 2). The robot responded by engaging in joint reference by orienting to the block. The subject was then asked to re-acquire the robot's attention (image 3) and reach for the toy (image 4-5). The robot observed this reach, classified it as an intent to approach or acquire the toy, and reached for the toy in response (image 5). The subject returns her attention to the robot and the robot engages her in mutual regard (image 6). Approximately 1.5 seconds elapsed between the images shown here.

- The distance between the actor and target is non-increasing.

When these two criteria are satisfied, an intentional relationship between the actor and the target is recorded. Similarly, an avoidance intention is recorded when:

- The distance between the actor and target is non-decreasing.
- The angle of the approach vector is maintained between 135 degrees and 225 degrees.

These criteria enforce that an avoidance relationship is assigned only when the actor is actively moving away from the target. Note that this assignment system allows for a single actor or object to maintain multiple intentional relationships (he wants x and y, he fears x and y). Figure 11-1 shows two examples of these attributions. In the left image, a person reaching toward a toy block is described as an attraction relation between the hand and the block. The right image in figure 11-1 shows the person withdrawing their hand from the block, an action which generates an avoidance event.

These intentional attributions can also be used to drive more complex behaviors. Because the intentional relationship is grounded to the perceptual properties of the location, color, and size of both the actor and the target, additional behavioral criteria can easily be applied. Furthermore, since each intentional relation is also grounded to the past history of movement for both agent and target, behavior can be selected based on the complete actions of the actor and the target. Figure 11-2 shows one example of an implemented behavioral connection between intentional relationships and a socially-engaging behavior. In this example, the robot was performing a set of behaviors including attempting to engage in joint reference (described in the previous chapter) and attempting to reach out toward the target of any observed attraction relationship. In this way, the robot acts "selfish" by attempting to obtain any object that is interesting enough for another person to desire (and approach) while not bothering to attempt to acquire objects that are only the object of visual attention. The interactions in this experiment were semi-scripted in that the experimenter asked the subject to perform a sequence of methods for engaging the robot socially. First, the subject was asked to get the robot's attention (image 1). The subject was then asked to turn and look at the toy block (image 2). The robot detected the change in head pose, which resulted in an increased saliency of the block, which caused a saccade and head orientation to the block. The subject was again asked to obtain the robot's attention (image 3). As a result, the robot saccaded and oriented again to the subject's face. Finally, the subject was asked to reach for the block (images 4-5). The robot observed the movement of the subject's left hand as an animate trajectory. ID detected an approach relationship between the moving hand and the block. The target of the approach relationship (the block) became the target of a reaching gesture (image 5). The subject then returned her attention to the robot. In response, the robot oriented to her and stopped acting on the intentional relationship (image 6). This example demonstrates the type of behavioral effects that intentional attribution can generate.

There are many possible extensions to this implementation of ID. Obviously, a richer set of intentional states and behavioral responses would enrich the system. Applications of intentional relationships as feedback mechanisms for social learning could also be explored. For example, it might be an adaptive strategy for a robot to learn to avoid classes of objects that humans avoid (such as pits or fire). This approach would also address many issues of generalization that have yet to be addressed in this context. Intentional states might also serve as an interesting starting point for implementations of episodic memory. Two of the critical problems in building systems that can remember what happened to them are in selecting the small set of data to be remembered from a very large input stream of data and in building a representational structure that can support this learning. The intentionality detector may provide a basis for addressing both of these problems; the intentional relationship provides the basis of a representational structure and also serves to pull out salient events as they occur.

Chapter 12

Toward a Theory of Mind

> *Isn't it curious that infants find social goals easier to accomplish than physical goals, while adults find the social goals more difficult? One way to explain this is to say that the presence of helpful people simplifies the infant's social world – since because of them, simpler actions solve harder problems ... How do children start on the path toward distinguishing between psychological and physical relationships?* – Minsky (1988, p. 297)

In the previous chapters, a novel model of the development of theory of mind was introduced, implemented on a humanoid robot, and evaluated in the context of social learning. This final chapter serves both to summarize the significant contributions of this implementation and to look beyond the current implementation toward a more complete theory of mind. We will consider the types of additional competencies that would be required to allow a machine to solve simple theory of mind tasks (such as the Smarties task or the Sally-Anne task described in chapter 3) and evaluate how the current implementation could support these extensions. We will also consider the implications that this existence proof provides in terms of the development of theory of mind abilities in human children

12.1 Summary of Significant Contributions

Based on the models of Baron-Cohen (1995) and Leslie (1994), we have proposed a hybrid model of the foundational skills for a theory of mind. This model, which we have called the *embodied theory of mind*, grounds concepts that have traditionally been thought to be high-level cognitive properties (such as animacy and intent) to low-level perceptual properties. All aspects of the model were implemented on a complex humanoid robot to operate in natural environments and at interactive rates. The implemented model featured the following components:

- An attentional mechanism which combined low-level feature detectors (such as color saturation, motion, and skin color filters) with high-level motivational influences to select regions of interest.

- A “theory of body” module which determined whether an object was animate or inanimate based on a set of naive physical laws that operated solely on the spatial and temporal properties of the object’s movement.

- An active sensorimotor system that detected faces at a large variety of scales using a color pre-filter and two shape-based metrics. This system also identified three features (the two eyes and the mouth) and used those features to determine the orientation of the person’s head. This information on the attentional state of the observed person was then used to engage in joint reference behaviors, directing the robot’s attention to the same object that the person was considering.

- A simple mechanism for detecting the basic intentional states of approach/desire and avoidance/fear. These classifications were determined by considering pairs of trajectories and allowing attributions of intent to only be applied to animate agents.

Individual components were evaluated by comparison with human judgments on similar problems and the complete system was evaluated in the context of social learning. A basic mimicry behavior was implemented by mapping a visual trajectory to a movement trajectory for one of Cog’s arms. Both the mimicry behavior and behaviors that generated an attentional reference (pointing and head orientation) were made socially relevant by limiting responses to animate trajectories, by acting on objects that became salient through joint reference, and by acting on objects that were involved in an intentional relationship. This set of simple behaviors made a first step toward constructing a system that can use natural human social cues to learn from a naive instructor.

12.2 Future Extensions

There are many obvious improvements that could be made to the sensory and motor control systems. Better perceptual systems for detecting the direction of eye gaze, for analyzing articulated motion, and for interpreting pointing gestures would be welcome additions. More complex motor behaviors such as object manipulation, whole-body movements, and coordinated manipulations between the two arms would allow for a wider range of social responses and richer interaction. It is easy to articulate these abilities and the effects that they might have on the system. Additional cognitive abilities would also increase the believability and usability of the system, but are more difficult to integrate into the existing behavioral architecture. For example, consider episodic memory which is the ability to represent, store, and recall events that have been experienced. The addition of episodic memory to this system would certainly provide new behavioral possibilities, but the exact effects of this inclusion and the ways in which episodic memory would affect the individual components of the system would need to be considered carefully. Beyond the obvious applications for learning sequences of movements, episodic memory might also allow the robot to use previously imitated behaviors to bias perception or to influence attentional selection. Other cognitive abilities might allow existing behaviors to be enhanced. The addition of a representation of objects that allowed identification across multiple viewpoints and representation outside the current visual field of view would permit joint reference behaviors to proceed to the representational stage outlined by Butterworth (1991).

While many additional cognitive capacities would add obvious value to this system, this implementation was designed to support three additional skills which were not part of this implementation: the attribution of belief and knowledge, more complex social learning mechanisms, and systems that show a basic level of self-awareness.

12.2.1 Attribution of Belief and Knowledge

The culmination of both the model of Leslie (1994) and the model of Baron-Cohen (1995) is the ability to attribute belief states to other individuals. The ability to represent the knowledge that another individual has is both a critical developmental waypoint and a useful skill in predicting behavior. No existing computational systems can currently pass any of the false belief tasks (such as the Smarties task or the Sally-Anne task) for real-world stimuli.[1] The core of Baron-Cohen's model is the thesis that the same foundational skills of joint reference, attribution of animacy, and inference of intent are the critical precursors in building a system that can evaluate these propositions about the internal knowledge states of other agents.

In order to apply knowledge states to other people, the robot must first have an internal representation of these knowledge states for itself. This representation might result from episodic memory, from object representations, or from a combination of both. For example, a simple form of episodic memory might connect sequences of judgments of intent or animacy with underlying perceptual properties. If the robot always saw red objects violate ToBY's energy expert (and thus be classified as animate) and then become the actor in an intentional relationship of attraction, an episodic memory system might generalize this into a rule about the behavior of red agents. When a person was present and was attending to the red object while the robot was making these generalizations, the agent's representation would be extended to include this rule about the behavior of red agents. Similar attributions might also be possible with a more complex object representation. For example, the robot might learn the preferences of other individuals by observing the characteristics of the objects that they tended to approach.

The architecture presented here could be extended to allow attribution of knowledge and belief states based on shared perceptual experiences. Informally, the system would operate by applying the rule that "if he sees what I see, then he knows what I know." Whenever an individual and the robot are engaged in a joint reference state, the robot would apply the same representational knowledge structures that it was combining at a given time to the other agent. From this point, it would also be

[1] It is trivial to build a system that can perform the appropriate inferences if given high-level abstractions of the perceptual data and the rules to be applied.

possible to begin to have the robot selectively predict the missing pieces of knowledge in another individual and attempt to supply that information by directing their attention to the relevant stimuli. If the robot was watching something slowly sneak up on you, it might point in that direction in an attempt to draw your attention to the potential danger. In this way, the robot might become not only a more active and involved student but also a very rudimentary instructor. Note that this differs drastically from a simple behavioral rule of the form "if you see something sneaking up, then you should point at it." The theory of mind abilities would be much more flexible, robust, and adaptive than these hard-coded rules. Ideally, since the attribution of knowledge states depends only on the referential process and not on the content of the information being conveyed, the robot could apply this information sharing technique without a deep understanding of the content of the message.

12.2.2 More Complex Social Learning Mechanisms

The mimicry mechanism described here is a relatively simple social learning mechanism. To achieve the grand challenge of building a machine that can acquire new tasks, skills, and knowledge from a naive instructor, much more complex social learning techniques will be required. In many ways, the foundations laid by this model of a theory of mind contribute to this challenge.

One obvious extension would be to move from mimicry to imitation. The simple example presented in chapter 11 in which the robot performed a reach for an object whenever it detected another agent's intent to approach that object was a first bridge between mimicry and imitation. The robot responded not to the raw movement, but rather to the intention of the agent (and the target of that intent). To make any real claims about systems that are capable of imitation, a much richer set of potential behavioral responses, triggering percepts, and intentional categories would be required. However, the same basic architecture could be used to support these more complex components. Intentional acts should serve as both the behavioral selection mechanism and as the guiding force that links a behavior to objects and events in the world. In this way, the robot would move beyond acting just as a mirror of the human's surface behavior and become a more autonomous and believable agent within the world. The transformation from mimicry to imitation signals a fundamental change in the way that any system, biological or artificial, represents other individuals.

A further step in developing socially competent systems would combine a goal-directed system with social exploration behaviors to create a robot that actively attempts to learn something. Imagine that the robot observed a person obtaining some desired object (perhaps a special toy) from within a locked box. If the robot desired that toy (for whatever reason), it might engage in behaviors that either directly or indirectly resulted in the human opening the box. Perhaps the robot would point and gesture to the box. Perhaps it would attempt to open the box, look despondent, and engage the human in an appeal for assistance. Perhaps it would engage in a more deceptive attempt to obtain the toy by distracting the person once the box was opened and then grabbing the toy for itself. The combinations of social learning with representations of knowledge and intent leads to issues of deception in robotic systems in the same ways that it does for humans (Wimmer & Perner, 1983; LaFreniere, 1988), other primates (Woodruff & Premack, 1979; Savage-Rumbaugh & McDonald, 1988; Hauser, 1992; Whiten & Byrne, 1988; Byrne & Whiten, 1991), and other animals (Ristau, 1991*a*,*b*). The development of robotic models capable of modeling knowledge states will be forced to address these issues.

12.2.3 Self-Aware Systems

One other area that has been closely related to theory of mind research in biology is the study of self perception and self awareness. Studies on human children have focused on how a child learns to distinguish itself from other people and from inanimate objects in the environment, how the child learns to use self-referential words and phrases, and how children learn to solve false belief tasks (Povinelli & Simon, 1998). Studies of animals have focused on self-recognition as an indicator of the cognitive capability of various species. Gallup (1970) first discussed self-recognition in non-human primates by observing their reactions to a mirror after placing an odorless, colored mark on the foreheads of chimpanzees. These animals had been exposed previously to mirrors and allowed to acclimate to these strange devices. The critical question involved in this study was whether the animal would reach for its own forehead or for the colored spot on the "other" animal that appeared in the mirror. If the animal reached for its own forehead, it would necessarily have developed some internal representation that the mirror allowed it to view its own body and not some other animal that just happened to act similarly. While both the exact findings and the usefulness of the method have been questioned, many researchers have used this as a critical thought experiment in developing other tests of self-recognition. Gallup's task is still being used, and recent research indicates that some other animals including monkeys (Hauser et al., 1995) and dolphins (Reiss & Marino, 2001) might recognize the creatures in the mirror as themselves.

In considering this question for a humanoid robot, a somewhat different set of research questions emerges based on the foundational skills of theory of mind. One of the basic functions of these skills is to distinguish between inanimate objects and agents which can be engaged socially. However, the current implementations classify the robot's own movements as animate. If a mirror is placed in front of the robot, it is perfectly content to engage itself indefinitely.[2] One could argue that a robot is unlikely to encounter many mirrors in some environments, however, the same problems arise whenever the robot happens to look down at its own body or whenever its arm happens to move up into its field of view. At some level, we would like the robot to recognize the contingency of the motor signals it is sending and the perceptual stimuli that co-occur. A system with a basic form of self-awareness would be able

[2] Yes, this test has actually been performed. The robot views the movement of its own head and arm as animate and will attempt to mimic that movement. Because the robot's movement never exactly matches its perception, the gesture being performed gradually changes over time.

to distinguish between perceived movements that were a result of its own body and those that originated from other sources.

This simple form of self-awareness might also be extended to guide social behaviors to further refine the class of socially-receptive stimuli. For example, it might be inappropriate for the robot to engage a pre-recorded television program while remaining appropriate for the robot to become a participant in a video conference call. One possible way of distinguishing between these different levels of social engagement is to play a simple imitation game.[3] In this game, the robot would alternate between attempting to mimic what it saw another individual doing and performing its own actions in an attempt to get the person to imitate it. By looking at the quality of the imitative responses in these two situations, the robot can start to distinguish objects that are interesting to engage socially. A pre-recorded television program might give the robot good quality actions for it to imitate, but would be very unlikely to imitate the actions that the robot initiates. Conversely, a mirror would be extremely good at following the movements that the robot initiates, while providing very little spontaneous movement for the robot to imitate. Static objects or objects that had physical morphologies that differed greatly from the robot would be poor both at providing quality movements to imitate and at responding to the robot's actions. Finally, those people interested in engaging the robot socially, whether they are physically in the same room or projected on a television screen, would be good at both phases of this imitation game.

12.3 Implications to Models of Human Behavior

Although no claims have been made that this implementation reflects the kinds of processing that occurs in either humans or other animals, systems like this one represent a new kind of tool in the evaluation and testing of human cognitive models (Adams et al., 2000; Webb, 2001). In particular, this implementation is an existence proof for building joint reference behaviors without an explicit, encapsulated module. The implementation has also demonstrated a useful addition to Wolfe's Guided Search model by incorporating both habituation effects and the effects of joint reference. Furthermore, the implemented system gives an example of how to perceptually ground animacy and intentionality judgments in real perceptual streams.

In a more general sense, robotic systems represent the natural next step in cognitive and behavioral modeling. Just as computer simulations presented researchers with the ability to make predictions for models that were difficult or impossible to see with only pen and paper, embodied models will provide predictions for models that rely upon complex interactions with the world that are difficult to simulate. Models of human social functioning rely upon the complex interplay between multiple people and the environment; performing simulations that can represent the wide variability and range of human responses is an extremely daunting task. By building systems

[3] I am indebted to Kerstin Dautenhahn and Cynthia Breazeal for assistance in developing this example.

that exist in the real world and interact directly with people, the difficult simulation problems disappear. Obviously, building real-world systems introduces an entirely new set of difficulties, but these problems often reflect deep underlying challenges that both biological and artificial systems must address.

An implemented robotic model also has benefits over direct experimentation on human subjects. Accurate testing and validation of these models through controlled, repeatable experiments can be performed. Slight experimental variations can be used to isolate and evaluate single factors (whether environmental or internal) independent of many of the confounds that affect normal behavioral observations. Experiments can also be repeated with nearly identical conditions to allow for easy validation. Further, internal model structures can be manipulated to observe the quantitative and qualitative effects on behavior. A robotic model can be subjected to testing that is potentially hazardous, costly, or unethical to conduct on humans; the "boundary conditions" of the models can be explored by testing alternative learning and environmental conditions. A robotic implementation may also be preferable to simulation studies or a theoretical analysis because the robot can interact freely with the same environmental conditions. Especially for models of social interaction, theoretical studies or simulations have great difficulty in accurately representing the complexities of agents in the environment. Finally, a robotic model can be used to suggest and evaluate potential educational strategies before applying them to human subjects.

12.4 Implications to Social Robotics

Technological devices rapidly become frustrating when they do not meet our expectations on how they can be used. In user interface design, the mapping between the user's goal and the actions that must be performed to achieve that goal should be as simple and obvious as possible. Rather than requiring users to learn some esoteric and exact programming language or interface, more and more systems are beginning to use the natural social interfaces that people use with each other. People continuously use this extremely rich and complex communication mechanism with seemingly little effort. The desire to have technologies that are responsive to these same social cues will continue to drive the development of systems that do what we want, not necessarily what we say. As artificial intelligence technology and robotics become more and more a part of our daily lives, these lessons will be even more important.

Theory of mind skills will be central to any technology that interacts with people. People attribute beliefs, goals, and desires to other agents so readily and naturally that it is extremely difficult for them to interact without using these skills. They will expect technology to do the same.

Bibliography

Adams, B. (2000), Meso: A Virtual Musculature for Humanoid Motor Control, Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Adams, B., Breazeal, C., Brooks, R. & Scassellati, B. (2000), 'Humanoid Robotics: A New Kind of Tool', *IEEE Intelligent Systems* **15**(4), 25–31.

Ashby, W. R. (1960), *Design for a Brain*, second edn, Chapman and Hall, London, United Kingdom. The first edition was published in 1952.

Aslin, R. N. (1987), Visual and Auditory Development in Infancy., *in* J. D. Osofksy, ed., 'Handbook of infant development, 2nd Ed.', Wiley, New York.

Atran, S. (1998), 'Folk Biology and the Anthropology of Science: Cognitive Universals and Cultural Particulars', *Behavioral and Brain Sciences* **21**(4), 547–569.

Baldwin, D. A. (1991), 'Infants' contribution to the achievement of joint reference', *Child Development* **62**, 875–890.

Ballard, D., Hayhoe, M. & Pelz, J. (1995), 'Memory representations in natural tasks', *Journal of Cognitive Neuroscience* **7**(1), 66–80.

Baluja, S. & Pomerleau, D. (1994), Non-Intrusive Gaze Tracking Using Artificial Neural Networks, Technical Report CMU-CS-94-102, Carnegie Mellon University.

Bar-Shalom, Y. & Formtann, T. E. (1988), *Tracking and Data Association*, Academic Press.

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

Baron-Cohen, S., Leslie, A. & Frith, U. (1985), 'Does the autistic child have a "theory of mind"?', *Cognition* **21**, 37–46.

Bernardino, A. & Santos-Victor, J. (1999), 'Binocular Visual Tracking: Integration of Perception and Control', *IEEE Transactions on Robotics and Automation* **15(6)**, 1937–1958.

Billard, A. & Dautenhahn, K. (1998), 'Grounding communication in autonomous robots: an experimental study.', *Robotics and Autonomous Systems.* **1–2**(24), 71–81.

Bingham, G. P., Schmidt, R. C. & Rosenblum, L. D. (1995), 'Dynamics and the Orientation of Kinematic Forms in Visual Event Recognition', *Journal of Experimental Psychology: Human Perception and Performance* **21**(6), 1473–1493.

Blythe, P. W., Todd, P. M. & Miller, G. F. (1999), How Motion Reveals Intention: Categorizing Social Interactions, *in* G. Gigerenzer & P. M. Todd, eds, 'Simple Heuristics that Make Us Smart', Oxford University Press, pp. 257–285.

Breazeal, C. (2000), Sociable Machines: Expressive Social Exchange Between Humans and Robots, PhD thesis, Massachusetts Institute of Technology.

Breazeal, C. & Scassellati, B. (1999), A context-dependent attention system for a social robot, *in* '1999 International Joint Conference on Artificial Intelligence'.

Breazeal, C. & Scassellati, B. (2001*a*), 'Infant-like Social Interactions between a Robot and a Human Caretaker', *Adaptive Behavior*. To appear.

Breazeal, C. & Scassellati, B. (2001*b*), Issues in Building Robots that Imitate People, *in* K. Dautenhahn & C. L. Nehaniv, eds, 'Imitation in Animals and Artifacts', MIT Press, chapter 14. To appear.

Breazeal, C., Edsinger, A., Fitzpatrick, P. & Scassellati, B. (2000*a*), Social Constraints on Animate Vision, *in* 'Proceedings of the First International IEEE/RSJ Conference on Humanoid Robotics'.

Breazeal, C., Edsinger, A., Fitzpatrick, P., Scassellati, B. & Varchavskaia, P. (2000*b*), 'Social constraints on animate vision', *IEEE Intelligent Systems* **15**(4), 32–37.

Brooks, R. A. (1986), 'A Robust Layered Control System for a Mobile Robot', *IEEE Journal of Robotics and Automation* **2**, 14–23.

Brooks, R. A. (1991*a*), Intelligence Without Reason, *in* 'Proceedings of the 1991 International Joint Conference on Artificial Intelligence', pp. 569–595.

Brooks, R. A. (1991*b*), 'Intelligence Without Representation', *Artificial Intelligence Journal* **47**, 139–160. Originally appeared as MIT AI Memo 899 in May 1986.

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

Brooks, R. A., Breazeal, C., Marjanović, M., Scassellati, B. & Williamson, M. M. (1999), The Cog Project: Building a Humanoid Robot, *in* C. L. Nehaniv, ed., 'Computation for Metaphors, Analogy and Agents', Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, Springer-Verlag.

Brooks, R. A., Breazeal (Ferrell), C., Irie, R., Kemp, C. C., Marjanović, M., Scassellati, B. & Williamson, M. M. (1998), Alternative Essences of Intelligence, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

Burghardt, G. (1990), Cognitive ethology and critical anthropomorphism: A snake with two heads and hog-nosed snakes that play dead, *in* C. Ristau, ed., 'Cognitive Ethology: The Minds of Other Animals', Erlbaum.

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.

Butterworth, G. (2000), 'Joint Attention is Based on the Facts of Embodiment and Not on a Theory of Mind', `http://www.warwick.ac.uk/fac/soc/Philosophy/consciousness/abstracts/Butterworth.html`.

Byrne, R. & Whiten, A. (1991), Computation and mindreading in primate tactical deception, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Byrne, R. & Whiten, A., eds (1988), *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.*, Oxford University Press.

Byrne, W. (1999), 'Imitation without intentionality. Using string parsing to copy the organization of behavior.', *Animal Cognition* **2**, 63–72.

Cannon, S. & Zahalak, G. I. (1982), 'The mechanical behavior of active human skeletal muscle in small oscillations', *Journal of Biomechanics* **15**, 111–121.

Carey, S. (1995), On the origin of causal understanding, *in* D. Sperber, D. Premack & A. J. Premack, eds, 'Causal cognition: A multidisciplinary debate', Symposia of the Fyssen Foundation. Fyssen Symposium, 6th January 1993., Oxford University Press, New York, pp. 268–308.

Carey, S. (1999), Sources of conceptual change, *in* E. K. Scholnick, K. Nelson, S. A. Gelman & P. H. Miller, eds, 'Conceptual Development: Piaget's Legacy', Lawrence Erlbaum Associates, pp. 293–326.

Chang, Y. L. & Aggarwal, J. K. (1991), 3D structure reconstruction from an ego motion sequence using statistical estimation and detection theory., *in* 'IEEE Workshop on Visual Motion', pp. 268–273.

Chaput, H. H. & Cohen, L. M. (2001), A Model of Infant Causal Perception and its Development, *in* 'Proceedings of the 2001 Cognitive Science Society Meeting'. In press.

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

Cheney, D. L. & Seyfarth, R. M. (1991), Reading Minds or Reading Behavior? Tests for a Theory of Mind in Monkeys, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Churchland, P., Ramachandran, V. & Sejnowski, T. (1994), A Critique of Pure Vision, *in* C. Koch & J. Davis, eds, 'Large-Scale Neuronal Theories of the Brain', MIT Press.

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

Cohen, L. B. & Amsel, G. (1998), 'Precursors to infants' perception of the causality of a simple event', *Infant Behavior and Develoment* **21**(4), 713–732.

Cohen, M. & Massaro, D. (1990), 'Synthesis of visible speech', *Behaviour Research Methods, Intruments and Computers* **22**(2), 260–263.

Coombs, D. & Brown, C. (1993), 'Real-Time Binocular Smooth Pursuit', *International Journal of Computer Vision* **11**(2), 147–164.

Cox, I. J. & Hingorani, S. L. (1996), 'An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **18**(2), 138–150.

Damasio, A. R. (1994), *Descartes' Error*, G.P. Putnam's Sons, New York.

Darrell, T., Gordon, G., Harville, M. & Woodfill, J. (1998*a*), Integrated person tracking using stereo, color, and pattern detection., *in* 'Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR-98)', pp. 601–609.

Darrell, T., Gordon, G., Harville, M. & Woodfill, J. (1998*b*), Integrated person tracking using stereo, color, and pattern detection, *in* 'Proceedings IEEE Conf. on Computer Vision and Pattern Recognition', pp. 601–608.

Dautenhahn, K. (1995), 'Getting to know each other–Artificial social intelligence for autonomous robots', *Robotics and Autonomous Systems* **16**(2–4), 333–356.

Dautenhahn, K. (1997), Ants Don't Have Friends – Thoughts on Socially Intelligent Agents, Technical report, AAAI Technical Report FS 97-02.

Dautenhahn, K. & Nehaniv, C. L., eds (2001), *Imitation in Animals and Artifacts*, MIT Press. To appear.

Demiris, J. & Hayes, G. (1999), Active and passive routes to imitation., *in* 'Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts', Edinburgh, pp. 81–87.

Dennett, D. C. (1987), *The Intentional Stance*, MIT Press.

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* 'The Development and Neural Bases of Higher Cognitive Functions', Vol. 608, New York Academy of Sciences, pp. 637–676.

Driscoll, J. A., Peters, II, R. A. & Cave, K. R. (1998), A visual attention network for a humanoid robot, *in* 'Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-98)'.

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

Edsinger, A. (2001), A Gestural Language for a Humanoid Robot, Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Edsinger, A., O'Reilly, U.-M. & Breazeal, C. (2000), Personality Through Faces for Humanoid Robots, *in* 'IEEE International Workshop on Robot and Human Communication (ROMAN-2000)'.

Emerson, R. W. (1860), *Behavior from The Conduct of Life (essays)*, Ticknor and Fields.

Fagan, J. F. (1976), 'Infants' recognition of invariant features of faces', *Child Development* **47**, 627–638.

Feigenbaum, E. A. & Feldman, J., eds (1963), *Computers and Thought*, McGraw-Hill, New York.

Ferrell, C. & Kemp, C. (1996), An Ontogenetic Perspective to Scaling Sensorimotor Intelligence, *in* 'Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium', AAAI Press.

Fodor, J. (1992), 'A theory of the child's theory of mind', *Cognition* **44**, 283–296.

Freud, S. (1962), *The ego and the id.*, Norton, New York.

Frith, C. D. & Frith, U. (1999), 'Interacting Minds – A Biological Basis', *Science* **286**, 1692–1695.

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

Galef, Jr., B. G. (1988), Imitation in animals: History, definitions, and interpretation of data from the psychological laboratory, *in* T. Zentall & B. G. Galef, eds, 'Social learning: Psychological and biological perspectives', Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 3–28.

Gallup, G. G. (1970), 'Chimpanzees: self-recognition', *Science* **167**, 86–87.

Gaussier, P., Moga, S., Banquet, J. P. & Quoy, M. (1998), 'From perception-action loops to imitation processes: A bottom-up approach of learning by imitation', *Applied Artificial Intelligence Journal* **12**(7–8), 701–729.

Gazzaniga, M. S. & LeDoux, J. E. (1978), *The Integrated Mind*, Plenum Press, New York.

Gee, A. H. & Cipolla, R. (1994), Determining the Gaze of Faces in Images, Technical Report CUED/F-INFENG/TR-174, University of Cambridge.

Gelman, R. (1990), 'First principles organize attention to and learning about relevant data: number and the animate-inanimate distinction as examples', *Cognitive Science* **14**, 79–106.

Gelman, R., Spelke, E. & Meck, E. (1983), What preschoolers know about animate and inanimate objects, *in* D. Rodgers & J. Sloboda, eds, 'The Acquisition of Symbolic Skills', Plenum Press, New York.

Gleitman, H. (1991), *Psychology*, third edn, W.W. Norton & Co., chapter 8.

Goldberg, M. E., Eggers, H. M. & Gouras, P. (1992), The Ocular Motor System, *in* E. R. Kandel, J. H. Schwartz & T. M. Jessell, eds, 'Principles of Neural Science', 3rd edn, Appleton and Lange.

Gomez, J. C. (1991), Visual Behavior as a Window for Reading the Mind of Others in Primates, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Graf, H. P., Chen, T., Petajan, E. & Cosatto, E. (1996), Locating Faces and Facial Parts, Technical Report TR-96.4.1, AT&T Bell Laboratories.

Greene, P. H. (1982), 'Why is it easy to control your arms?', *Journal of Motor Behavior* **14**(4), 260–286.

Griggs, R. A. & Cox, J. R. (1982), 'The elusive thematics material effect in Wason's selection task.', *British Journal of Psychology* **73**, 407–420.

Hashimoto, S. (1998), Humanoid Robots in Waseda University - Hadaly-2 and WABIAN, *in* 'IARP First International Workshop on Humanoid and Human Friendly Robotics', Tsukuba, Japan.

Hauser, M. D. (1992), 'Costs of Deception: Cheaters are Punished in Rhesus Monkeys', *Proc. Natl. Acad. Sci.* **89**, 12137–12139.

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

Hauser, M., Kralik, J., Botto-Mahan, C., Garrett, M. & Oser, J. (1995), 'Self-recognition in primates: Phylogeny and the salience of species-typical features', *Proc. Natl. Acad. Sci.* **92**, 10811–10814.

Hayes, G. M. & Demiris, J. (1994), A Robot Controller Using Learning by Imitation, *in* 'Proceedings 2nd International Symposium on Intelligent Robotic Systems', Grenoble, France, pp. 198–204.

Heider, F. & Simmel, M. (1944), 'An experimental study of apparent behavior.', *American Journal of Psychology* **57**, 243–259.

Heinzmann, J. & Zelinsky, A. (1997), Robust Real-Time Face Tracking and Gesture Recognition, *in* '1997 International Joint Conference on Artificial Intelligence', Vol. 2, pp. 1525–1530.

Herman, L. (2001), Vocal, social, and self-imitation by bottlenosed dolphins, *in* K. Dautenhahn & C. L. Nehaniv, eds, 'Imitation in Animals and Artifacts', MIT Press. To appear.

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

Horn, B. K. P. (1986), *Robot Vision*, MIT Press.

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

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.

Itti, L., Koch, C. & Niebur, E. (1998), 'A Model of Saliency-Based Visual Attention for Rapid Scene Analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **20**(11), 1254–1259.

Johnson, M. H. (1993), Constraints on Cortical Plasticity, *in* M. H. Johnson, ed., 'Brain Development and Cognition: A Reader', Blackwell, Oxford, pp. 703–721.

Jones, M. J. & Viola, P. (2001), Robust Real-Time Object Detection, Technical Report CRL 2001/01, Compaq Cambridge Research Laboratory.

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

Julesz, B. & Bergen, J. R. (1983), 'Textons, the Fundamental Elements in Preattentive Vision and Perception of Textures', *The Bell System Technical Journal* **62**(6), 1619–1645.

Julesz, B. & Krose, B. (1988), 'Features and spatial filters', *Nature* **333**, 302–303.

Kanwisher, N. & Downing, P. (1998), 'Separating the Wheat from the Chaff', *Science* **282**, 57–58.

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.

Keil, F. C. (1995), The growth of causal understandings of natural kinds, *in* D. Sperber, D. Premack & A. J. Premack, eds, 'Causal cognition: A multidisciplinary debate', Symposia of the Fyssen Foundation. Fyssen Symposium, 6th January 1993., Oxford University Press, New York, pp. 234–267.

Kirpatrick, S., Gelatt, Jr., C. & Vecchi, M. (1993), 'Optimization by Simulated Annealing', *Science* **220**, 671–680.

Knudsen, E. I. & Knudsen, P. F. (1985), 'Vision Guides the Adjustment of Auditory Localization in Young Barn Owls', *Science* **230**, 545–548.

Kozima, H. (1998), Attention-sharing and behavior-sharing in human-robot communication, *in* 'IEEE International Workshop on Robot and Human Communication (ROMAN-98, Takamatsu)', pp. 9–14.

Kuniyoshi, Y., Inaba, M. & Inoue, H. (1994), 'Learning by watching: Extracting reusable task knowledge from visual observation of human performance.', *IEEE Transactions on Robotics and Automation* **10**(6), 799–822.

Kuniyoshi, Y., Kita, N., Sugimoto, K., Nakamura, S. & Suehiro, T. (1995), A Foveated Wide Angle Lens for Active Vision, *in* 'Proc. IEEE Int. Conf. Robotics and Automation'.

LaFreniere, P. J. (1988), The ontongeny of tactical deception in humans, *in* R. Byrne & A. Whiten, eds, 'Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.', Oxford University Press.

Lakoff, G. (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago, Illinois.

Leslie, A. M. (1982), 'The perception of causality in infants', *Perception* **11**, 173–186.

Leslie, A. M. (1984), 'Spatiotemporal continuity and the perception of causality in infants', *Perception* **13**, 287–305.

Leslie, A. M. (1994), ToMM, ToBY, and Agency: Core architecture and domain specificity, *in* L. A. Hirschfeld & S. A. Gelman, eds, 'Mapping the Mind: Domain specificity in cognition and culture', Cambridge University Press, pp. 119–148.

Leslie, A. M. & Keeble, S. (1987), 'Do six-month old infants perceive causality?', *Cognition* **25**, 265–188.

Lisberger, S. G. & Sejnowski, T. J. (1992), 'Motor learning in a recurrent network model based on the vestibulo-ocular reflex', *Nature* **260**, 159–161.

Lorenz, K. (1973), *Foundations of Ethology*, Springer-Verlag, New York, NY.

Loy, G., Holden, E. J. & Owens, R. (2000), A 3D head tracker for an automated lipreading system, *in* 'Proceedings of the Australian Conference on Robotics and Automation (ACRA 2000)', pp. 37–43.

Lund, N. J. & Duchan, J. F. (1983), *Assessing Children's Language in Naturalistic Contexts*, Prentice-Hall, Englewood Cliffs, NJ, chapter Cognitive Precursors to Language Acquisition.

Mack, A. & Rock, I. (1998), *Inattentional Blindness*, MIT Press.

MacKay, W. A., Crammond, D. J., Kwan, H. C. & Murphy, J. T. (1986), 'Measurements of human forearm posture viscoelasticity', *Journal of Biomechanics* **19**, 231–238.

Marjanović, M. (1995), Learning Functional Maps Between Sensorimotor Systems on a Humanoid Robot, Master's thesis, MIT Department of Electrical Engineering and Computer Science.

Marjanović, M. (2001), Teach a Robot to Fish... A Thesis Proposal, Technical report, Massachusetts Institute of Technology. Available from `http://www.ai.mit.edu/people/maddog`.

Marjanović, M. J., Scassellati, B. & Williamson, M. M. (1996), Self-Taught Visually-Guided Pointing for a Humanoid Robot, *in* 'From Animals to Animats: Proceedings of 1996 Society of Adaptive Behavior', Society of Adaptive Behavior, Cape Cod, Massachusetts, pp. 35–44.

Matarić, M. J., Williamson, M. M., Demiris, J. & Mohan, A. (1998), Behaviour-Based Primitives for Articulated Control, *in* R. Pfiefer, B. Blumberg, J.-A. Meyer & S. W. Wilson, eds, 'Fifth International Conference on Simulation of Adaptive Behavior', The MIT Press, Cambridge, MA, pp. 165–170.

Maurer, T. & von der Malsburg, C. (1996), Tracking and Learning Graphs and Pose on Image Sequences of Faces, *in* 'Proc. 2nd Int. Conf. on Automatic Face- and Gesture-Recognition', IEEE Press, pp. 176–181.

Meltzoff, A. & Moore, K. (1994), 'Imitation, memory, and the representation of persons.', *Infant Behavior and Development* **17**, 83–99.

Meltzoff, A. N. (1995), 'Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children', *Developmental Psychology* **31**, 838–850.

Michotte, A. (1962), *The perception of causality*, Methuen, Andover, MA.

Minsky, M. (1988), *The Society of Mind*, Simon and Schuster.

Minsky, M. & Papert, S. (1970), 'Draft of a proposal to ARPA for research on artificial intelligence at MIT, 1970-71'.

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

Morimoto, C., Koons, D., Amir, A. & Flickner, M. (1998), Pupil Detection and Tracking Using Multiple Light Sources, Technical Report RJ10117, IBM Research Journal.

Mundy, P. & Sigman, M. (1989), 'The theoretical implications of joint attention deficits in autism', *Development and Psychopathology* **1**, 173–183.

Mussa-Ivaldi, F. A., Hogan, N. & Bizzi, E. (1985), 'Neural, Mechanical, and Geometric Factors Subserving Arm Posture in humans', *Journal of Neuroscience* **5**(10), 2732–2743.

Nakayama, K. & Silverman, G. H. (1986), 'Serial and Parallel Processing of Visual Feature Conjunctions', *Nature* **320**, 264–265.

Nehaniv, C. & Dautenhahn, K. (1998), Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications, *in* J. Demiris & A. Birk, eds, 'Learning Robots: An Interdisciplinary Approach', World Scientific Press.

Newell, A. & Simon, H. (1961), GPS, a program that simulates thought, *in* H. Billing, ed., 'Lernende Automaten', R. Oldenbourg, Munich, Germany, pp. 109–124. Reprinted in (Feigenbaum and Feldman, 1963, pp.279–293).

Norman, D. A. (1990), *The Design of Everyday Things*, Doubleday.

Nothdurft, H. C. (1993), 'The role of features in preattentive vision: Comparison of orientation, motion and color cues', *Vision Research* **33**, 1937–1958.

Nummenmaa, T. (1964), *The Language of the Face*, Vol. 9 of *University of Jyvaskyla Studies in Education, Psychology and Social Research.* Reported in Baron-Cohen (1995).

Panerai, F. & Sandini, G. (1998), 'Oculo-Motor Stabilization Reflexes: Integration of Inertial and Visual Information', *Neural Networks* **11**(7/8), 1191–1204.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA.

Perner, J. & Lang, B. (1999), 'Development of theory of mind and executive control', *Trends in Cognitive Sciences.*

Perner, J., Frith, U., Leslie, A. M. & Leekam, S. (1989), 'Exploration of the autistic child's theory of mind: knowledge, belief, and communication', *Child Development* **60**, 689–700.

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

Povinelli, D. J. & Simon, B. B. (1998), 'Young children's understanding of briefly versus extremely delayed images of the self: Emergence of the autobiographical stance.', *Developmental Psychology* **34**(1), 188–194.

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.

Premack, D. (1990), 'The infant's theory of self-propelled objects', *Cognition* **36**, 1–16.

Premack, D. & Woodruff, G. (1978), 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences* **4**, 515–526.

Reeves, B. & Nass, C. (1996), *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press.

Reid, D. B. (1979), 'An algorithm for tracking multiple targets', *IEEE Transactions on Automated Control* **24**(6), 843–854.

Reiss, D. & Marino, L. (2001), 'Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence', *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5937–5942.

Rensink, R., O'Regan, J. & Clark, J. (1997), 'To See or Not to See: The Need for Attention to Perceive Changes in Scenes', *Psychological Science* **8**, 368–373.

Ristau, C. (1991*a*), Attention, purposes, and deception in birds, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

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

Rosales, R. & Sclaroff, S. (1999), Inferring Body Pose without Tracking Body Parts, Technical Report 1999-017, Boston University.

Rosenschein, S. J. & Kaelbling, L. P. (1986), The Synthesis of Machines with Provable Epistemic Properties, *in* J. Halpern, ed., 'Proceedings of the Conference on Theoretical Aspects of Reasoning about Knowledge', Morgan Kaufmann Publishers, Los Altos, California, pp. 83–98.

Rougeaux, S. & Kuniyoshi, Y. (1997), Velocity and Disparity Cues for Robust Real-Time Binocular Tracking, *in* 'IEEE Proc. Computer Vision and Pattern Recognition', pp. 1–6.

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

Savage-Rumbaugh, S. & McDonald, K. (1988), Deception and social manipulation in symbol-using apes, *in* R. Byrne & A. Whiten, eds, 'Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.', Oxford University Press.

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. (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 American Association of Artificial Intelligence (AAAI-98)', pp. 969–976.

Scassellati, B. (1999*a*), Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot, *in* C. L. Nehaniv, ed., 'Computation for Metaphors, Analogy and Agents', Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, Springer-Verlag.

Scassellati, B. (1999*b*), Knowing what to imitate and knowing when you succeed, *in* 'Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts', pp. 105–113. April 6-9.

Schaal, S. (1997), Robot learning from demonstration, *in* D. H. Fisher, Jr., ed., 'International Conference on Machine Learning (ICML-97)', Morgan Kaufmann, San Francisco, CA, pp. 12–20.

Schaal, S. (1999), 'Is imitation learning the route to humanoid robots?', *Trends in Cognitive Sciences* **3**(6), 233–242.

Scholl, B. J. & Tremoulet, P. D. (2000), 'Perceptual causality and animacy', *Trends in Cognitive Sciences* **4**(8), 299–309.

Searle, J. R. (1986), *Minds, Brain and Science*, Harvard University 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.

Sinha, P. (1997), Personal Communication, August, 1997.

Steels, L. (1996), Emergent Adaptive Lexicons, *in* 'Proceedings of the fourth international conference on simulation of adaptive behavior', Cape Cod, MA, pp. 562–567.

Stroop, J. (1935), 'Studies of interference in serial verbal reactions', *Journal of Experimental Psychology* **18**, 643–62.

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

Terzopoulous, D. & Waters, K. (1991), Techniques for realistic facial modeling and animation, *in* M. Magnenat-Thalmann & D. Thalmann, eds, 'Computer Animation '91', Springer-Verlag.

Thayer, S. (1977), 'Children's detection of on-face and off-face gazes', *Developmental Psychology* **13**, 673–674.

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

Tomasi, C. & Kanade, T. (1992), 'Shape and motion from image streams under orthography: a factorization method', *International Journal of Computer Vision* **9**(2), 137–154.

Treisman, A. (1985), 'Preattentive processing in vision', *Computer Vision, Graphics, and Image Processing* **31**, 156–177.

Tsotsos, J. K. (1995), 'Behaviorist intelligence and the scaling problem', *Artificial Intelligence* **75**(2), 135–160.

Turing, A. M. (1950), 'Computing Machinery and Intelligence', *Mind* **49**, 433–460.

Turk, M. & Pentland, A. (1991), 'Eigenfaces for recognition', *Journal of Cognitive Neuroscience.*

van der Spiegel, J., Kreider, G., Claeys, C., Debusschere, I., Sandini, G., Dario, P., Fantini, F., Belluti, P. & Soncini, G. (1989), A foveated retina-like sensor using CCD technology, *in* C. Mead & M. Ismail, eds, 'Analog VLSI implementation of neural systems', Kluwer Academic Publishers, pp. 189–212.

Wason, P. C. (1966), Reasoning, *in* B. M. Foss, ed., 'New Horizons in Psychology', Vol. 1, Penguin Books, Harmondsworth, England, pp. 135–51.

Webb, B. (2001), 'Can robots make good models of biological behaviour?', *Behavioral and Brain Sciences.*

Weiskrantz, L. (1986), *Blindsight: A Case Study and Implications*, Vol. 12 of *Oxford Psychology Series*, Clarendon Press, Oxford.

Wertheimer, M. (1961), 'Psychomotor coordination of auditory and visual space at birth', *Science* **134**, 1692.

Whiten, A. & Byrne, R. W. (1988), The manipulation of attention in primate tactical deception, *in* R. Byrne & A. Whiten, eds, 'Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.', Oxford University Press.

Whiten, A. & Byrne, R. W., eds (1997), *Machiavellian Intelligence II: Extensions and Evaluations*, Cambridge University Press.

Whiten, A. & Ham, R. (1992), 'On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research', *Advances in the Study of Behaviour* **21**, 239–283.

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

Williamson, M. M. (1999), Robot Arm Control Exploiting Natural Dynamics, PhD thesis, Massachusetts Institute of Technology.

Wimmer, H. & Perner, J. (1983), 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception', *Cognition* **13**, 103–128.

Wolfe, J. M. (1994), 'Guided Search 2.0: A revised model of visual search', *Psychonomic Bulletin & Review* **1**(2), 202–238.

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

Woodruff, G. & Premack, D. (1979), 'Intentional communication in the chimpanzee: The development of deception', *Cognition* **7**, 333–362.

Yeshurun, Y. & Schwartz, E. L. (1989), 'Cepstral Filtering on a Columnar Image Architecture: A Fast Algorithm for Binocular Stereo Segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7), 759–767.

Zajac, F. E. (1989), 'Muscle and tendon:Properties, models, scaling, and application to biomechanics and motor control', *CRC Critical Reviews of Biomedical Engineering* **17**(4), 359–411.