The Cog Project: Building a Humanoid Robot

Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanović,
Brian Scassellati, Matthew M. Williamson

MIT Artificial Intelligence Lab
545 Technology Square
Cambridge MA 02139, USA
{brooks,cynthia,maddog,scaz,matt}@ai.mit.edu
http://www.ai.mit.edu/projects/cog/

Abstract. To explore issues of developmental structure, physical embodiment, integration of multiple sensory and motor systems, and social interaction, we have constructed an upper-torso humanoid robot called Cog. The robot has twenty-one degrees of freedom and a variety of sensory systems, including visual, auditory, vestibular, kinesthetic, and tactile senses. This chapter gives a background on the methodology that we have used in our investigations, highlights the research issues that have been raised during this project, and provides a summary of both the current state of the project and our long-term goals. We report on a variety of implemented visual-motor routines (smooth-pursuit tracking, saccades, binocular vergence, and vestibular-ocular and opto-kinetic reflexes), orientation behaviors, motor control techniques, and social behaviors (pointing to a visual target, recognizing joint attention through face and eye finding, imitation of head nods, and regulating interaction through expressive feedback). We further outline a number of areas for future research that will be necessary to build a complete embodied system.

1 Introduction

Building an android, an autonomous robot with humanoid form and human-like abilities, has been both a recurring theme in science fiction and a "Holy Grail" for the Artificial Intelligence community. In the summer of 1993, our group began the construction of a humanoid robot. This research project has two goals: an engineering goal of building a prototype general purpose flexible and dextrous autonomous robot and a scientific goal of understanding human cognition (Brooks & Stein 1994).

Recently, many other research groups have begun to construct integrated humanoid robots (Hirai, Hirose, Haikawa & Takenaka 1998, Kanehiro, Mizuuchi, Koyasako, Kakiuchi, Inaba & Inoue 1998, Takanishi, Hirano & Sato 1998, Morita, Shibuya & Sugano 1998). There are now conferences devoted solely to humanoid systems, such as the International Symposium on Humanoid Robots (HURO) which was first hosted by Waseda University in October of 1996, as well as sections of more broadly-based conferences, including a recent session at the 1998

C. Nehaniv (Ed.): Computation for Metaphors, Analogy, and Agents, LNCS 1562, pp. 52-87, 1999.

IEEE International Conference on Robotics and Automation (ICRA-98) in Leuven, Belgium. There has also been a special issue of the Journal of the Robotics Society of Japan in October of 1997 devoted solely to humanoid robotics.

Research in humanoid robotics has uncovered a variety of new problems and a few solutions to classical problems in robotics, artificial intelligence, and control theory. This research draws upon work in developmental psychology, ethology, systems theory, philosophy, and linguistics, and through the process of implementing models and theories from these fields has raised interesting research issues. In this chapter, we review some of the methodology and results from the first five years of our humanoid robotics project.

Since the inception of our research program, we have developed a methodology that departs from the mainstream of AI research (Brooks, Breazeal (Ferrell), Irie, Kemp, Marjanović, Scassellati & Williamson 1998). Section 2 reviews some of the assumptions of classical AI that we have found lacking and concentrates on four aspects of a new methodology that have greatly influenced our research program: developmental structure, physical embodiment, integration of multiple sensory and motor systems, and social interaction. In section 3, we describe the current hardware and software environments of our upper-torso humanoid robot, including twenty-one mechanical degrees of freedom, a variety of sensory systems, and a heterogeneous distributed computation system. Section 4 focuses on some of the long-term research issues that members of our group are currently investigating, and Section 5 describes some of the current tasks and behaviors that our robot is capable of performing. We conclude in Section 6 with a few of the open problems that have yet to be addressed.

2 Methodology

In recent years, AI research has begun to move away from the assumptions of classical AI: monolithic internal models, monolithic control, and general purpose processing. However, these concepts are still prevalent in much current work and are deeply ingrained in many architectures for intelligent systems. For example, in the recent AAAI-97 proceedings, one sees a continuing interest in planning (Littman 1997, Hauskrecht 1997, Boutilier & Brafman 1997, Blythe & Veloso 1997, Brafman 1997) and representation (McCain & Turner 1997, Costello 1997, Lobo, Mendez & Taylor 1997), which build on these assumptions.

Previously, we have presented a methodology that differs significantly from the standard assumptions of both classical and neo-classical artificial intelligence (Brooks et al. 1998). Our alternative methodology is based on evidence from cognitive science and neuroscience which focus on four alternative attributes which we believe are critical attributes of human intelligence: developmental organization, social interaction, embodiment and physical coupling, and multimodal integration.

In this section, we summarize some of the evidence that has led us to abandon those assumptions about intelligence that classical AI continues to uphold. We then briefly review the alternative methodology that we have been using in constructing humanoid robotic systems.

2.1 False Assumptions about Human Intelligence

In studying human intelligence, three common conceptual errors often occur: reliance on monolithic internal models, on monolithic control, and on general purpose processing. These and other errors primarily derive from naive models based on subjective observation and introspection, and biases from common computational metaphors (mathematical logic, Von Neumann architectures, etc.)(Brooks 1991*a*, Brooks 1991*b*). A modern understanding of cognitive science and neuroscience refutes these assumptions.

Humans have no full monolithic internal models. There is evidence that in normal tasks humans tend to minimize their internal representation of the world. Ballard, Hayhoe & Pelz (1995) have shown that in performing a complex task, like building a copy of a display of blocks, humans do not build an internal model of the entire visible scene. By changing the display while subjects were looking away, 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.

There is also evidence that there are multiple internal representations, which are not mutually consistent. For example, in the phenomena of blindsight, cortically blind patients can discriminate different visual stimuli, but report seeing nothing (Weiskrantz 1986). This inconsistency would not be a feature of a single central model of visual space.

These experiments and many others like it, e.g. Rensink, O'Regan & Clark (1997) and Gazzaniga & LeDoux (1978), convincingly demonstrate that humans do not construct a full, monolithic model of the environment. Instead humans tend to only represent what is immediately relevant from the environment, and those representations do not have full access to one another.

Humans have 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). As an experimental treatment for severe epilepsy in these patients, the corpus callosum (the main structure connecting the two hemispheres of the brain) was surgically cut. The patients are surprisingly normal after the operation, but with deficits that are revealed by presenting different information to either side of the (now unconnected) brain. Since each hemisphere controls one side of the body, the experimenters can probe the behavior of each hemisphere independently (for example, by observing the subject picking up an object appropriate to the scene that they had viewed). In one example, a snow scene was presented to the right hemisphere and the leg of a chicken to the left. 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 independently acted appropriately, but one side falsely explained the choice of the other. This suggests that there are multiple independent control systems, rather than a single monolithic one.

Humans are 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). When presented with a list of words written in a variety of colors, performance in a color recognition and articulation task is dependent on the semantic content of the words; the task is very difficult if names of colors are printed in non-corresponding colors. This experiment demonstrates the specialized nature of human computational processes and interactions.

Even in the areas 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. However, with additional context—labeling the cards such that they were understandable as names and ages—subjects could easily solve exactly the same problem.

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).

2.2 Essences of Human Intelligence

In an attempt to simplify the problem of building complex intelligent systems, classical AI approaches tended to ignore or avoid many aspects of human intelligence (Minsky & Papert 1970). We believe that many of these discarded elements are essential to human intelligence. Our methodology exploits four central aspects of human intelligence: development, social interaction, physical interaction and integration. 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. We believe that not only are these four themes critical to the understanding of human intelligence but also they actually simplify the problem of creating human-like intelligence.

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 perform incrementally more difficult tasks in more complex environments en route to the adult state. 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.

Development is an *incremental* process. Behaviors and learned skills that have already been mastered prepare and enable the acquisition of more advanced behaviors by providing subskills and knowledge that can be re-used, by placing simplifying constraints on the acquisition, and by minimizing new information that must be acquired. For example, 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 this progression, infants in later phases consistently demonstrate more sophisticated reaching strategies to retrieve a toy in more challenging scenarios. As the infant's reaching competency develops, later stages incrementally improve upon the competency afforded by the previous stages. Within our group, Marjanović, Scassellati & Williamson (1996) applied a similar bootstrapping technique to enable the robot to learn to point to a visual target. Scassellati (1996) has discussed how a humanoid robot might acquire basic social competencies through this sort of developmental methodology. Other examples of developmental learning that we have explored can be found in (Ferrell 1996, Scassellati 1998*b*).

By gradually increasing the complexity of the required task, a developmental process optimizes learning. For example, infants are born with low acuity vision which simplifies the visual input they must process. 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 in which the acuity of both sensory and motor systems are gradually increased significantly reduces the difficulty of the learning problem (Thelen & Smith 1994). The caregiver also acts to gradually increase the task complexity by structuring and controlling the complexity of the environment. 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 scale autonomously (Ferrell & Kemp 1996, Scassellati 1998*b*).

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. This reliance on social contact is so integrated into our species that it is hard to imagine a completely asocial human; developmental disorders that effect social development, such as autism and Asperger's syndrome, are extremely debilitating and can have far-reaching consequences (Cohen & Volkmar 1997). 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. Our research program has investigated social interaction both as a means for bootstrapping and as an instance of developmental progression.

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 infant to experience the end or outcome before she is cognitively or physically able of seeking and attaining it for herself (Wood, Bruner & Ross 1976). We are currently engaged in work studying bootstrapping new behaviors from social interactions (Breazeal & Scassellati 1998, Breazeal & Velasquez 1998).

The social skills required to make use of scaffolding are complex. Infants acquire these social skills through a developmental progression (Hobson 1993). One of the earliest precursors is the ability to share attention with the caregiver. This ability can take many forms, from the recognition of a pointing gesture to maintaining eye contact (see chapter in this volume by Scassellati). In our work, we have also examined social interaction from this developmental perspective, building systems that can recognize and respond to joint attention by finding faces and eyes (Scassellati 1998*c*) and imitating head nods of the caregiver (Scassellati 1998*d*).

Embodiment and Physical Coupling: Perhaps the most obvious, and most overlooked, aspect of human intelligence is that it is embodied. 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 both to allow humans to interact socially with the robot in a natural way and to provide similar task constraints.

The direct physical coupling between action and perception reduces the need for an intermediary representation. For an embodied system, internal representations can be ultimately grounded in sensory-motor interactions with the world (Lakoff 1987). Our systems are physically coupled with the world and operate directly in that world without any explicit representations of it (Brooks 1986, Brooks 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.[1]

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. Embodied systems can also perform some complex tasks in relatively simple ways by exploiting the properties of the complete system. 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. We have implemented one such scheme using self-adaptive oscillators to drive the joints of the robot's arm (Williamson 1998*a*, Williamson 1998*b*).

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. For example, Churchland, Ramachandran & Sejnowski (1994) demonstrated an example of how audition can cause illusory visual motion. Vision can cause auditory illusions too, such as the McGurk effect (Cohen & Massaro 1990). These studies demonstrate that sensory modalities cannot be treated independently.

Sensory integration can simplify the computation necessary for a given task. Attempting to perform the task using only one modality is sometimes awkward and computationally intensive. Utilizing the complementary nature of separate modalities can result in a reduction of overall computation. We have implemented several mechanisms on Cog that use multimodal integration to aid in increasing performance or developing competencies. For example, Peskin & Scassellati (1997) implemented a system that stabilized images from a moving camera using vestibular feedback.

By integrating different sensory modalities we can exploit the multimodal nature of stimuli to facilitate learning. For example, objects that make noise often move. This correlation can be exploited to facilitate perception. Wertheimer (1961) has shown that vision and audition interact from birth; even ten-minute-old children will turn their eyes toward an auditory cue. This interaction between the senses continues to develop; visual stimuli greatly affect the development of sound localization (Knudsen & Knudsen 1985). In our work, Irie (1997) built an auditory system that utilizes visual information to train auditory localization. This work highlights not only the development of sensory integration, but also

[1] This was the fundamental approach taken by Ashby (1960) contemporaneously with the development of early AI.

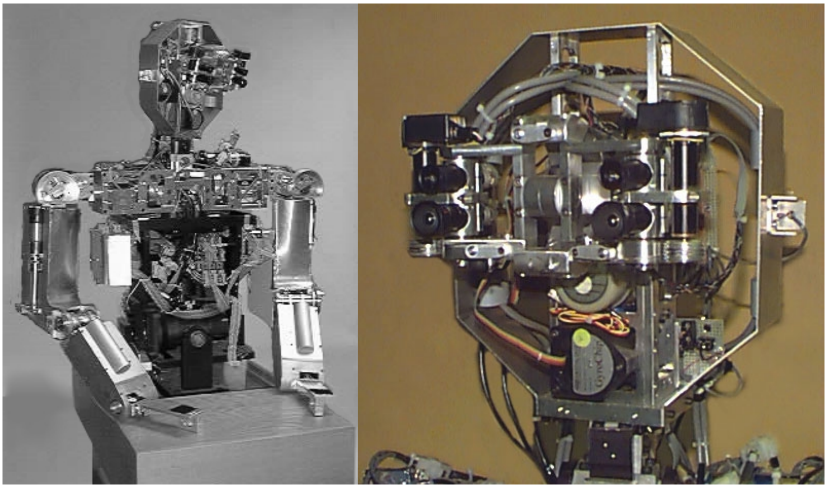

Fig. 1. Cog, an upper-torso humanoid robot. Cog has twenty-one degrees of freedom to approximate human movement, and a variety of sensory systems that approximate human senses, including visual, vestibular, auditory, and tactile senses.

the simplification of computational requirements that can be obtained through integration.

3 Hardware

In pursuing the methodology outlined in the previous section, we have constructed an upper-torso humanoid robot called Cog (see Figure 1). This section describes the computational, perceptual, and motor systems that have been implemented on Cog as well as the development platforms that have been constructed to test additional hardware and software components.

3.1 Computational System

The computational control for Cog is a heterogeneous network of many different processors types operating at different levels in the control hierarchy, ranging from small microcontrollers for joint-level control to digital signal processor (DSP) networks for audio and visual preprocessing.

Cog's "brain" has undergone a series of revisions. The original was a network of 16 MHz Motorola 68332 microcontrollers on custom-built boards, connected through dual-port RAM. Each of these nodes ran L, a multithreading subset of Common Lisp. The current core is a network of 200 MHz industrial PC computers running the QNX real-time operating system and connected by 100VG ethernet. The network currently contains 4 nodes, but can be expanded at will by plugging new nodes into the network hub. QNX provides transparent and fault-tolerant interprocess communication over the network. The PC backplanes provide ample room for installing commercial or custom I/O boards and controller cards. The "old" and "new" brains can inter-operate, communicating via a custom-built shared memory ISA interface card.

Video and audio preprocessing is performed by a separate network of Texas Instruments C40 digital signal processors which communicate via the proprietary C40 communications port interface. The network includes C40-based framegrabbers, display boards, and audio I/O ports. The processors relay data to the core processor network via ISA and PCI interface cards.

Each joint on the robot has a dedicated local motor controller, a custom-built board with a Motorola HC11 microcontroller, which processes encoder and analog inputs, performs servo calculations, and drives the motor via pulse-width modulation. For the arms, the microcontroller generates a virtual spring behavior at 1kHz, based on torque feedback from strain gauges in the joints.

3.2 Perceptual Systems

To obtain information about the environment, Cog has a variety of sensory systems including visual, vestibular, auditory, tactile, and kinesthetic senses.

Visual System: Cog's visual system is designed to mimic some of the capabilities of the human visual system, including binocularity and space-variant sensing (Scassellati 1998*a*). Each eye can rotate about an independent vertical axis (pan) and a coupled horizontal axis (tilt). To allow for both a wide field of view and high resolution vision, there are two grayscale cameras per eye, one which captures a wide-angle view of the periphery ($88.6°(V) \times 115.8°(H)$ field of view) and one which captures a narrow-angle view of the central (foveal) area ($18.4°(V) \times 24.4°(H)$ field of view with the same resolution). Each camera produces an NTSC signal that is digitized by a frame grabber connected to the digital signal processor network.

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 has three rate gyroscopes mounted on orthogonal axes (corresponding to the semi-circular canals) and two linear accelerometers (corresponding to the otolith organs). Each of these devices is mounted in the head of the robot, slightly below eye level. Analog signals from each of these sensors is amplified on-board the robot, and processed off-board by a commercial A/D converter attached to one of the PC brain nodes.

Auditory System: To provide auditory information, two omni-directional microphones were mounted on the head of the robot. To facilitate localization, crude pinnae were constructed around the microphones. Analog auditory signals are processed by a commercial A/D board that interfaces to the digital signal processor network.

Tactile System: We have begun experimenting with providing tactile feedback from the robot using resistive force sensors. Each sensor provides a measurement of the force applied to its sensing surface. As an initial experiment, we have mounted an 6×4 array of these sensors to the front of the robot's torso. The signals from these sensors are multiplexed through a single 6811 microcontroller, thus giving measurements of both force and position. A similar system has been used to mount tactile sensors on some of the hands that we have used with the robot.

Kinesthetic System: Feedback concerning the state of Cog's motor system is provided by a variety of sensors located at each joint. The eye axes utilize only the simplest form of feedback; each actuator has a single digital encoder which gives position information. The neck and torso joints have encoders, as well as motor current sensing (for crude torque feedback), temperature sensors on the motors and driver chips, and limit switches at the extremes of joint movement. The arms 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.

3.3 Motor Systems

Cog has a total of twenty-one mechanical degrees-of-freedom (DOF); two six DOF arms, a torso with a two degree-of-freedom (DOF) waist, a one DOF torso twist, a three DOF neck, and three DOF in the eyes.

Arms: Each arm is loosely based on the dimensions of a human arm with 6 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, Crammond, Kwan & Murphy 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

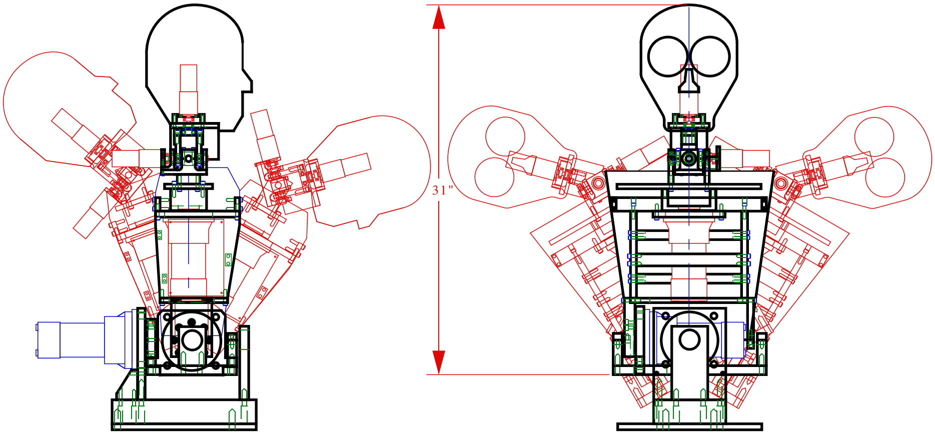

Fig. 2. Range of motion for the neck and torso. Not shown are the neck twist (180 degrees) and body twist (120 degrees)

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.

Neck and Torso: Cog's body has six degrees of freedom: the waist bends side-to-side and front-to-back, the "spine" can twist, and the neck tilts side-to-side, nods front-to-back, and twists left-to-right. Mechanical stops on the body and neck give a human-like range of motion, as shown in Figure 2 (Not shown are the neck twist (180 degrees) and body twist (120 degrees)).

3.4 Development Platforms

In addition to the humanoid robot, we have also built three development platforms, similar in mechanical design to Cog's head, with identical computational systems; the same code can be run on all platforms. These development platforms allow us to test and debug new behaviors before integrating them on Cog.

Vision Platform: The vision development platform (shown at the left of Figure 3) is a copy of Cog's active vision system. The development platform has identical degrees of freedom, similar design characteristics, and identical computational environment. The development platform differs from Cog's vision system in only three ways. First, to explore issues of color vision and saliency, the development platform has color cameras. Second, the mechanical design of the camera mounts

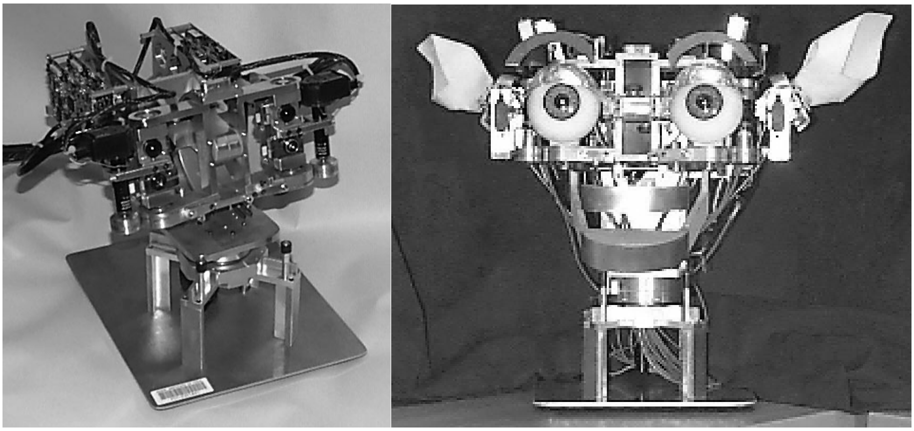

Fig. 3. Two of the vision development platforms used in this work. These desktop systems match the design of the Cog head and are used as development platforms for visual-motor routines. The system on the right has been modified to investigate how expressive facial gestures can regulate social learning.

has been modified for the specifications of the color cameras. Third, because the color cameras are significantly lighter than the grayscale cameras used on Cog, we were able to use smaller motors for the development platform while obtaining similar eye movement speeds. Additional details on the development platform design can be found in Scassellati (1998*a*).

Vision and Emotive Response Platform: To explore ideas in social interaction between robots and humans, we have constructed a platform with capabilities for emotive facial expressions (shown at the right of Figure 3). This system, called Kismet, consists of the active stereo vision system (described above) embellished with facial features for emotive expression. Currently, these facial features include 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), and a mouth (with one degree of freedom: open/close). The robot is able to show expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise (shown in Figure 4) which are easily interpreted by an untrained human observer.

A pair of Motorola 68332-based microcontrollers are also connected to the robot. One controller implements the motor system for driving the robot's facial motors. The second controller implements the motivational system (emotions and drives) and the behavior system. This node receives pre-processed perceptual information from the DSP network through a dual-ported RAM, and converts this information into a behavior-specific percept which is then fed into the rest of the behavior engine.

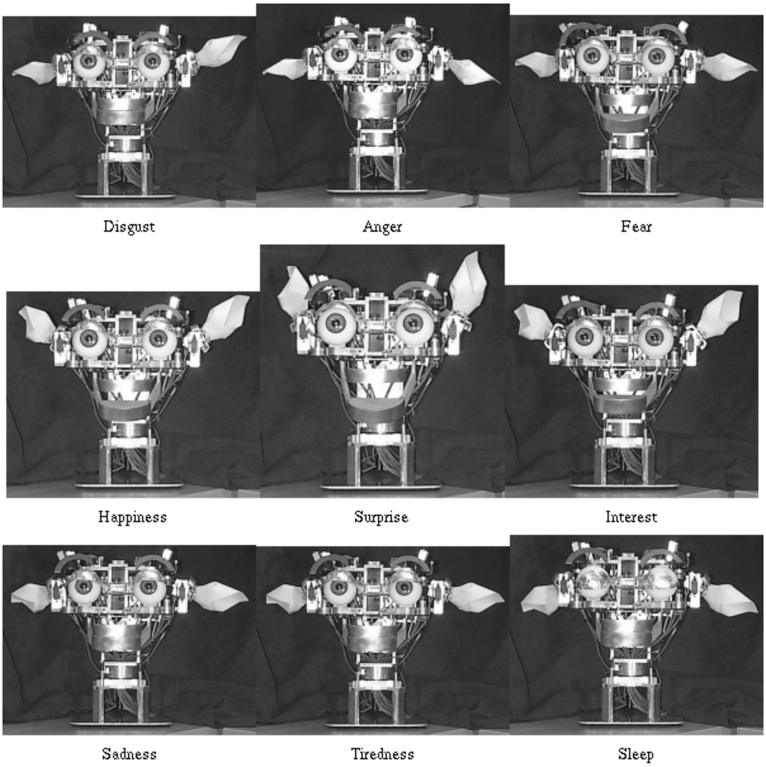

Fig. 4. Static extremes of Kismet's facial expressions. During operation, the 11 degrees-of-freedom for the ears, eyebrows, mouth, and eyelids vary continuously with the current emotional state of the robot.

Visual-Auditory Platform: A third development platform was constructed to investigate the relationships between vision and audition. The development platform has an auditory system similar to that used on Cog, with two microphones and a set of simplified pinnae. As a simplified visual system, a single color camera was mounted at the midline of the head.

4 Current Long-Term Projects

This section describes a few of the long-term research issues that our group is currently addressing. Although each project is still in progress, initial results from each of these areas will be presented in Section 5.

4.1 Joint Attention and Theory of Mind

One critical milestone in a child's development is the recognition of others as agents that have beliefs, desires, and perceptions that are independent of the child's own beliefs, desires, and perceptions. The ability to recognize what another person can see, the ability to know that another person maintains a false belief, and the ability to recognize that another person likes games that differ from those that the child enjoys are all part of this developmental chain. Further, the ability to recognize oneself in the mirror, the ability to ground words in perceptual experiences, and the skills involved in creative and imaginative play may also be related to this developmental advance. 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.

If we are to build a system that can recognize and produce these complex social behaviors, we must find a skill decomposition that maintains the complexity and richness of the behaviors represented while still remaining simple to implement and construct. Evidence from the development of these "theory of mind" skills in normal children, as well as the abnormal development seen in pervasive developmental disorders such as Asperger's syndrome and autism, demonstrate that a critical precursor is the ability to engage in joint attention (Baron-Cohen 1995, Frith 1990). Joint attention refers to those preverbal social behaviors that allow the infant to share with another person the experience of a third object (Wood et al. 1976). For example, the child might laugh and point to a toy, alternating between looking at the caregiver and the toy.

From a robotics standpoint, even the simplest of joint attention behaviors require the coordination of a large number of perceptual, sensory-motor, attentional, and cognitive processes. Our current research is the implementation of one possible skill decomposition that has received support from developmental psychology, neuroscience, and abnormal psychology, and is consistent with evidence from evolutionary studies of the development of joint attention behaviors. This decomposition is described in detail in the chapter by Scassellati, and requires many capabilities from our robotic system including basic eye motor skills, face and eye detection, determination of eye direction, gesture recognition, attentional systems that allow for social behavior selection at appropriate moments, emotive responses, arm motor control, image stabilization, and many others.

A robotic system that can recognize and engage in joint attention behaviors will allow for social interactions between the robot and humans that have previously not been possible. 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 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 models from natural systems 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.

4.2 Social Interaction between an Infant and a Caretaker

Other ongoing work focuses on altricial learning in a social context (Breazeal (Ferrell) 1998, Breazeal & Scassellati 1998, Breazeal & Velasquez 1998). By treating the robot as an altricial system whose learning is assisted and guided by the human caretaker, this approach exploits the environment and social interactions that are critical to infant development.

An infant's motivations (emotions, drives, and pain) play an important role in generating meaningful interactions with the caretaker (Bullowa 1979). The infant's emotional responses provide important cues which the caretaker uses to assess how to satiate the infant's drives, and how to carefully regulate the complexity of the interaction. The former is critical for the infant to learn how its actions influence the caretaker, and the latter is critical for establishing and maintaining a suitable learning environment for the infant. Similarly, the caretaker's emotive responses to the infant shape the continuing interaction and can guide the learning process.

An infant's motivations are vital to regulating social interactions with his mother (Kaye 1979). Soon after birth, an infant is able to display a wide variety of facial expressions (Trevarthen 1979). As such, he responds to events in the world with expressive cues that his mother can read, interpret, and act upon. She interprets them as indicators of his internal state (how he feels and why), and modifies her actions to promote his well being (Tronick, Als & Adamson 1979, Chappell & Sander 1979). For example, when he appears content she tends to maintain the current level of interaction, but when he appears disinterested she intensifies or changes the interaction to try to re-engage him. In this manner, the infant can regulate the intensity of interaction with his mother by displaying appropriate emotive and expressive cues.

An important function for a robot's motivational system is not only to establish appropriate interactions with the caretaker, but also to regulate their intensity so that the robot is neither overwhelmed nor under stimulated by them. When designed properly, the intensity of the robot's expressions provide appropriate cues for the caretaker to increase the intensity of the interaction, tone it down, or maintain it at the current level. By doing so, both parties can modify their own behavior and the behavior of the other to maintain the intensity of interaction that the robot requires.

The use of emotional expressions and gestures facilitates and biases learning during social exchanges. Parents take an active role in shaping and guiding how and what infants learn by means of *scaffolding.* As the word implies, the parent provides a supportive framework for the infant by manipulating the infant's interactions with the environment to foster novel abilities. The emotive cues the parent receives during social exchanges serve as feedback so the parent can adjust the nature and intensity of the structured learning episode to maintain a suitable learning environment where the infant is neither bored nor overwhelmed.

In addition, an infant's motivations and emotional displays are critical in establishing the context for learning shared meanings of communicative acts (Halliday 1975). An infant displays a wide assortment of emotive cues such as coos, smiles, waves, and kicks. At such an early age, the mother imparts a consistent meaning to her infant's expressive gestures and expressions, interpreting them as meaningful responses to her mothering and as indications of his internal state. Curiously, experiments by Kaye (1979) argue that the mother actually supplies most if not *all* the meaning to the exchange when the infant is so young. The infant does not know the significance his expressive acts have for his mother, nor how to use them to evoke specific responses from her. However, because the mother *assumes* her infant shares the same meanings for emotive acts, her consistency *allows* the infant to discover what sorts of activities on his part will get specific responses from her. Routine sequences of a predictable nature can be built up which serve as the basis of learning episodes (Newson 1979).

Combining these ideas one can design a robot that is biased to learn how its emotive acts influence the caretaker in order to satisfy its own drives. Toward this end, we endow the robot with a motivational system that works to maintain its drives within homeostatic bounds and motivates the robot to learn behaviors that satiate them. For our purposes, we further provide the robot with a set of emotive expressions that are easily interpreted by a naive observer as analogues of the types of emotive expressions that human infants display. This allows the caretaker to observe the robot's emotive expressions and interpret them as communicative acts. This establishes the requisite routine interactions for the robot to learn how its emotive acts influence the behavior of the caretaker, which ultimately serves to satiate the robot's own drives. By doing so, both parties can modify both their own behavior and the behavior of the other in order to maintain an interaction that the robot can learn from and use to satisfy its drives.

4.3 Dynamic Human-like Arm Motion

Another research goal is to build a system that can move with the speed, precision, dexterity, and grace of a human to physically interact with the world in human-like ways. Our current research focuses on control methods that exploit the natural dynamics of the robot to obtain flexible and robust motion without complex computation.

Control methods that exploit physical dynamics are not common in robotics. Traditional methods are often kinematically based, requiring accurate calibration of the robot's dimensions and mechanical properties. However, even for systems that utilize only a few degrees of freedom, kinematic solutions can be computationally expensive. For this reason, researchers have adopted a number of strategies to simplify the control problems by reducing the effects of system dynamics including careful calibration and intensive modeling (An, Atkeson & Hollerbach 1988), using lightweight robots with little dynamics (Salisbury, Townsend, Eberman & DiPietro 1988), or simply by moving slowly. Research emphasizing dynamic manipulation either exploits clever mechanical mechanisms which simplify control schemes (Schaal & Atkeson 1993, McGeer 1990) or results in computationally complex methods (Mason & Salisbury 1985).

Humans, however, exploit the mechanical characteristics of their bodies. For example, when humans swing their arms they choose comfortable frequencies which are close to the natural resonant frequencies of their limbs (Herr 1993, Hatsopoulos & Warren 1996). Similarly, when placed in a jumper, infants bounce at the natural frequency (Warren & Karrer 1984). Humans also exploit the active dynamics of their arm when throwing a ball (Rosenbaum et al. 1993) and the passive dynamics of their arm to allow stable interaction with objects (Mussa-Ivaldi, Hogan & Bizzi 1985). When learning new motions, both infants and adults quickly utilize the physical dynamics of their limbs (Thelen & Smith 1994, Schneider, Zernicke, Schmidt & Hart 1989).

On our robot, we have exploited the dynamics of the arms to perform a variety of tasks. The compliance of the arm allows both stable motion and safe interaction with objects. Local controllers at each joint are physically coupled through the mechanics of the arm, allowing these controllers to interact and produce coordinated motion such as swinging a pendulum, turning a crank, and playing with a slinky. Our initial experiments suggest that these solutions are very robust to perturbations, do not require accurate calibration or parameter tuning, and are computationally simple (Williamson 1998*a*, Williamson 1998*b*).

4.4 Multi-modal Coordination

Our group has developed many behaviors and skills for Cog, each involving one or two sensory and/or motor systems – i.e. face finding, crank turning, auditory localization. However, to be truly effective as an embodied robot, Cog requires a general mechanism for overall sensory-motor coordination, a facility for effectively combining skills or at least preventing them from interfering with each other.

A multi-modal coordination system will manifest itself in three different ways. First, for interactions between sensory systems, such a facility would provide a basis for the combination of several sensory inputs into a more robust and reliable view of the world. Second, interactions between motor systems produce synergisms — coactivation of motor systems not directly involved with a task but which prepare the robot for more effective execution overall. Third, for interactions between sensory and motor systems, this system would provide a method for "sensory tuning," in which adjusting physical properties of the robot can optimize the performance of a sensory system (foveation is a very basic example).

The foundation for such a general coordination mechanism rests on two modules: a system that incorporates intrinsic performance measures into sensorimotor processes, and a system for extracting correlations between sensorimotor events. Combined, these provide sufficient information for Cog to learn how its internal systems interact with each other. Unfortunately, finding this information is by no means trivial.

Performance measures are the most straightforward. For sensory processes, the performance is estimated by a confidence measure, probably based on a combination of repeatibility, error estimates, etc. Motor performance measurements would be based upon criteria such as power expenditure, fatigue measures, safety limits, and actuator accuracy.

Extracting correlations between sensorimotor events is more complex. The first step is segmentation, that is, determining what constitutes an "event" within a stream of proprioceptive data and/or motor commands. Segmentation algorithms and filters can be hard-coded (but only for the most rudimentary enumeration of sensing and actuating processes) or created adaptively. Adaptive segmentation creates and tunes filters based on how well they contribute to the correlation models. Segmentation is crucial because it reduces the amount of redundant information produced by confluent data streams. Any correlation routine must deal with both the combinatorial problem of looking for patterns between many different data sources and the problem of finding correlations between events with time delays.

A general system for multimodal coordination is too complex to implement all at once. We plan to start on a small scale, coordinating between two and five systems. The first goal is a mechanism for posture — to coordinate, fixate, and properly stiffen or relax torso, neck, and limbs for a variety of reaching and looking tasks. Posture is not merely a reflexive control; it has feed-forward components which require knowledge of impending tasks so that the robot can ready itself. A postural system being so reactive and pervasive, requires a significant amount of multi-modal integration.

5 Current Tasks

In pursuing the long-term projects outlined in the previous section, we have implemented many simple behaviors on our humanoid robot. This section briefly describes the tasks and behaviors that the robot is currently capable of performing. For brevity, many of the technical details and references to similar work have been excluded here, but are available from the original citations. In addition, video clips of Cog performing many of these tasks are available from `http://www.ai.mit.edu/projects/cog/`.

5.1 Visual-motor Routines

Human eye movements can be classified into five categories: three voluntary movements (saccades, smooth pursuit, and vergence) and two involuntary movements (the vestibulo-ocular reflex and the opto-kinetic response)(Goldberg, Eggers & Gouras 1992). We have implemented mechanical analogues of each of these eye motions.

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 $S : (x, e) \mapsto \Delta e$ which produces a change in eye motor position (Δe) given the current eye motor position (e) and the stimulus location in the image plane (x). To obtain accurate saccades without requiring an accurate model of the kinematics and optics, an unsupervised learning algorithm estimates the saccade function. This implementation can adapt to the non-linear optical and mechanical properties of the vision system. 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.

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. The central 7×7 patch of the initial 64×64 image is installed as the target image. In this instance, we use a very small image to reduce the computational load necessary to track non-artifact features of an object. For each successive image, the central 44×44 patch is correlated with the 7×7 target image. The best correlation value gives the location of the target within the new image, and the distance from the center of the visual field to that location gives the motion vector. The length of the motion vector is the pixel error. The motion vector is scaled by a constant (based on the time between iterations) and used as a velocity command to the motors. This system operates at 20 Hz. and can successfully track moving objects whose image projection changes slowly.

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 are critical for tasks such as accurate hand-eye coordination, figure-ground discrimination, and collision detection. Yamato (1998) built a system that performs binocular vergence and integrates the saccadic and smooth-pursuit systems described previously. Building on models of the development of binocularity in infants, Yamato used local correlations to identify matching targets in a foveal region in both eyes, moving the eyes to match the pixel locations of the targets in each eye. The system was also capable of smoothly responding to changes of targets after saccadic motions, and during smooth pursuit.

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). To mimic the human vestibular system, Cog has three rate gyroscopes mounted on orthogonal axis (corresponding to the semi-circular canals) and two linear accelerometers (corresponding to the otolith organs).

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, our computational load is manageable. The optic flow routine 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 that we have learned previously, we can obtain an estimate of the amount of eye motion we require to compensate for the visual displacement.

A simple VOR can be constructed by integrating the velocity signal from the rate gyroscopes, 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 (Peskin & Scassellati 1997). 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

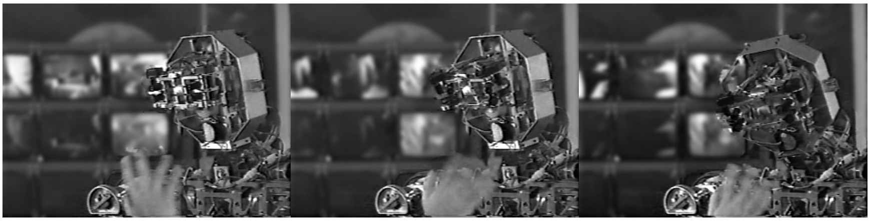

Fig. 5. Orientation to a salient stimulus. Once a salient stimulus (a moving hand) has been detected, the robot first saccades to that target and then orients the head and neck to that target.

drift from the VOR. We are currently working on implementing models of VOR and OKN coordination to allow both systems to operate simultaneously.

5.2 Eye/Neck Orientation

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. 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). To move the neck the appropriate distance, we must construct a mapping $N : (n, e) \mapsto \Delta n$ which produces a change in neck motor positions (Δn) given the current neck position (n) and the initial eye position (e). Because we are mapping motor positions to motor positions with axes that are roughly parallel, a simple linear mapping has sufficed: $\Delta n = (k\dot{e} - n)$ for some constant k.[2]

There are two possible mechanisms for counter-rotating the eyes while the neck is in motion: the vestibulo-ocular reflex or an efference copy signal of the neck motion. VOR can be used to compensate for neck motion without any additions necessary. Because the reflex uses gyroscope feedback to maintain the eye position, no communication between the neck motor controller and the eye motor controller is necessary. This can be desirable if there is limited bandwith between the processors responsible for neck and eye control. However, using VOR to compensate for neck motion can become unstable. Because the gyroscopes are mounted very close to the neck motors, motion of the neck can result in additional vibrational noise on the gyroscopes. However, since the neck motion is a voluntary movement, our system can utilize additional information in order to counter-rotate the eyes, much as humans do (Ghez 1992). An efference copy signal can be used to move the eye motors while the neck motors are moving. The neck motion signal can be scaled and sent to the eye motors to compensate for the neck motion. The scaling constant is simply $\frac{1}{k}$, where k is the same constant

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

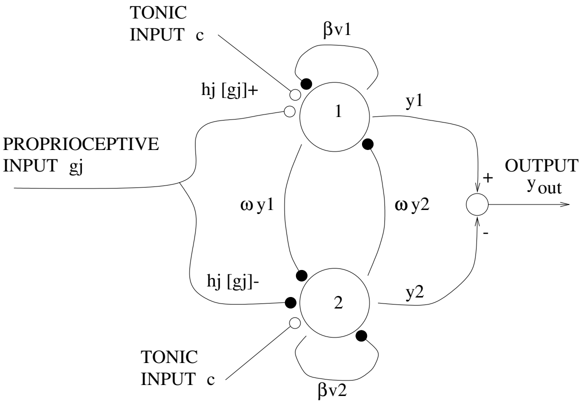

Fig. 6. Schematic of the oscillator. Black circles correspond to inhibitory connections, open circles to excitatory. The βv_i connections correspond to self-inhibition, and the ωy_i connections give the mutual inhibition. The positive and negative parts of the input g_j are weighted by the gain h_j before being applied to the neurons. The two outputs y_i are combined to give the oscillator output y_{out}.

that was used to determine Δn. Just as with the vestibulo-ocular reflex, the scaling constants can be obtained using controlled motion and feedback from the opto-kinetic nystigmus. Using efference copy with constants obtained from OKN training results in a stable system for neck orientation.

5.3 Dynamic Oscillator Motor Control

Neural oscillators have been used to generate repetitive arm motions. The coupling between a set of oscillators and the physical arm of the robot achieves many different tasks using the same software architecture and without explicit models of the arm or environment. The tasks include swinging pendulums at their resonant frequencies, turning cranks, and playing with a slinky.

Using a proportional-derivative control law, the torque at the ith joint can be described by:

$$u_i = k_i(\theta_{vi} - \theta_i) - b_i\dot{\theta}_i \tag{1}$$

where k_i is the stiffness of the joint, b_i the damping, θ_i the joint angle, and θ_{vi} the equilibrium point. By altering the stiffness and damping of the arm, the dynamical characteristics of the arm can be changed. The posture of the arm can be changed by altering the equilibrium points (Williamson 1996). This type of control preserves stability of motion. The elastic elements of the arm produce a system that is both compliant and shock resistant, allowing the arm to operate in unstructured environments.

Two simulated neurons with mutually inhibitory connections drive each arm joint, as shown in Figure 6. The neuron model describes the firing rate of a biological neuron with self-inhibition (Matsuoka 1985). The firing rate of each neuron is governed by the following equations:

$$\tau_1 \dot{x_1} = -x_1 - \beta v_1 - \omega \left[x_2\right]^+ - \Sigma_{j=1}^{j=n} h_j \left[g_j\right]^+ + c \tag{2}$$

$$\tau_2 \dot{v_1} = -v_1 + \left[x_1\right]^+ \tag{3}$$

$$\tau_1 \dot{x_2} = -x_2 - \beta v_2 - \omega \left[x_1\right]^+ - \Sigma_{j=1}^{j=n} h_j \left[g_j\right]^- + c \tag{4}$$

$$\tau_2 \dot{v_2} = -v_2 + \left[x_2\right]^+ \tag{5}$$

$$y_i = \left[x_i\right]^+ = max(x_i, 0) \tag{6}$$

$$y_{out} = y_1 - y_2 \tag{7}$$

where x_i is the firing rate, v_i is the self-inhibition of the neuron (modulated by the adaption constant β), and ω controls the mutual inhibition. The output of each neuron y_i is the positive portion of the firing rate, and the output of the whole oscillator is y_{out}. Any number of inputs g_j can be applied to the oscillator, including proprioceptive signals and signals from other neurons. Each input is scaled by a gain h_j and arranged to excite one neuron while inhibiting the other by applying the positive portion of the input ($\left[g_j\right]^+$) to one neuron and the negative portion to the other. The amplitude of the oscillation is proportional to the tonic excitation c. The speed and shape of the oscillator output are determined by the time constants τ_1 and τ_2. For stable oscillations, τ_1/τ_2 should be between 0.1 and 0.5. The stability and properties of this oscillator system and more complex networks of neurons are analyzed in Matsuoka (1985) and Matsuoka (1987).

The output of the oscillator y_{out} is connected to the equilibrium point θ_v. One neuron flexes the joint and the other extends it about a fixed posture θ_p, making the equilibrium point $\theta_v = y_{out} + \theta_p$. The inputs to the oscillators are either the force (τ) or the position (θ) of the joint.[3] The interaction of the oscillator dynamics and the physical dynamics of the arm form a tightly coupled dynamical system. Unlike a conventional control system, there is no "set-point" for the motion. The interaction of the two coupled dynamical systems determines the overall arm motion.

The oscillators have two properties which make them suitable for certain types of repetitive motions. First, they can entrain an input signal over a wide range of frequencies. In the entrained state, the oscillator provides an output at exactly the same frequency as the input, with a phase difference between input and output which depends on frequency. Second, the oscillator also becomes entrained very rapidly, typically within one cycle. Figure 7 shows the entrainment of an oscillator at the elbow joint as the shoulder of the robot is moved. The movement of the shoulder induces forces at the elbow which drive the elbow in synchrony with the shoulder.

[3] These signals in general have an offset (due to gravity loading, or other factors). When the positive and negative parts are extracted and applied to the oscillators, a low-pass filter is used to find and remove the DC component.

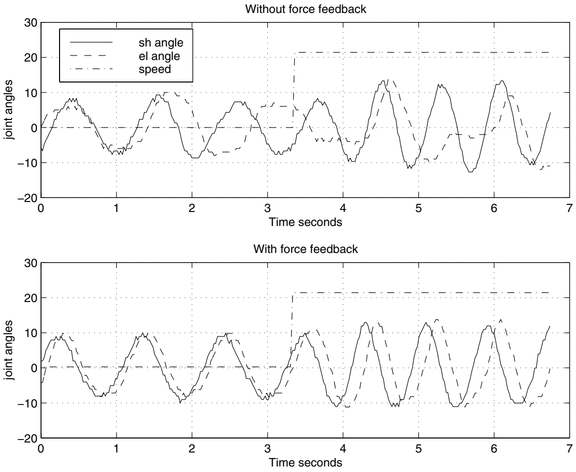

Fig. 7. Entrainment of an oscillator at the elbow as the shoulder is moved. The joints are connected only through the physical structure of the arm. Both plots show the angle of the shoulder (solid) and the elbow (dashed) as the speed of the shoulder is changed (speed parameter dash-dot). The top graph shows the response of the arm without proprioception, and the bottom with proprioception. Synchronization occurs only with the proprioceptive feedback.

Slinky: The entrainment property can be exploited to manipulate objects, such as a slinky. As the slinky is passed from hand to hand, the weight of the slinky is used to entrain oscillators at both elbow joints. The oscillators are completely independent, and unsynchronized, in software. With the slinky forming a physical connection between the two systems, the oscillators work in phase to produce the correct motion. The adaptive nature of the oscillators allows them to quickly recover from interruptions of motion and changes in speed. An example of the coordination is shown in Figure 8.

Cranks: The position constraint of a crank can also be used to coordinate the joints of the arm. If the arm is attached to the crank and some of the joints are moved, then the other joints are constrained by the crank. The oscillators can sense the motion, adapt, and settle into a stable crank turning motion.

In the future, we will explore issues of complex redundant actuation (such as multi-joint muscles), utilize optimization techniques to tune the parameters of the oscillator, produce whole-arm oscillations by connecting various joints into a single oscillator, and explore the use of postural primitives to move the set point of the oscillations.

5.4 Pointing to a Visual Target

We have implemented a pointing behavior which enables Cog to reach out its arm to point to a visual target (Marjanović et al. 1996). The behavior is learned

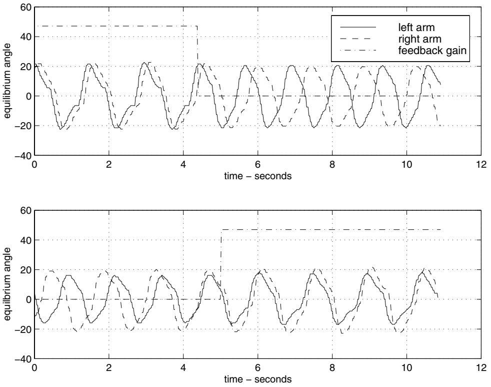

Fig. 8. The robot operating the slinky. Both plots show the outputs from the oscillators as the proprioception is turned on and off. With proprioception, the outputs are synchronized. Without proprioception, the oscillators move out of phase. The only connection between the oscillators is through the physical structure of the slinky.

over many repeated trials without human supervision, using gradient descent methods to train forward and inverse mappings between a visual parameter space and an arm position parameter space. This behavior uses a novel approach to arm control, and the learning bootstraps from prior knowledge contained within the saccade behavior (discussed in Section 5.1). As implemented, the behavior assumes that the robot's neck remains in a fixed position.

From an external perspective, the behavior is quite rudimentary. 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. Early reaches are inaccurate, and often in the wrong direction altogether, but after a few hours of practice the accuracy improves drastically.

The reaching algorithm involves an amalgam of several subsystems. A motion detection routine identifies a salient stimulus, which serves as a target for the saccade module. This foveation guarantees that the target is always at the center of the visual field; the coordinates of the target on the retina are always the center of the visual field, and the position of the target relative to the robot is wholly characterized by the gaze angle of the eyes (only two degrees of freedom). Once the target is foveated, the joint configuration necessary to point to that target is generated from the gaze angle of the eyes using a "ballistic map." This configuration is used by the arm controller to generate the reach.

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).

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 where Cog could have looked, but not how it should have moved the arm. To train the ballistic map, we also need a "forward map" — i.e. a forward kinematics function which gives the gaze angle of the hand in response to a commanded set of joint positions. 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.

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." A primitive is a fixed set joint angles, corresponding to a static position of the arm, placed at a corner 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 learning.

Unfortunately, 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.

5.5 Recognizing Joint Attention Through Face and Eye Finding

The first joint attention behaviors that infants engage in involve maintaining eye contact. To enable our robot to recognize and maintain eye contact, we have implemented a perceptual system capable of finding faces and eyes (Scassellati 1998*c*). The system first locates potential face locations in the peripheral image using a template-based matching algorithm developed by Sinha (1996). Once a potential face location has been identified, the robot saccades to that target using the saccade mapping S described earlier. The location of the face in peripheral image coordinates ($p_{(x,y)}$) is then mapped into foveal image coordinates ($f_{(x,y)}$) using a second learned mapping, the foveal map $F : p_{(x,y)} \mapsto f_{(x,y)}$. The location of the face within the peripheral image can then be used to extract the sub-image containing the eye for further processing.

This technique has been successful at locating and extracting sub-images that contain eyes under a variety of conditions and from many different individuals. Additional information on this task and its relevance to building systems that recognize joint attention can be found in the chapter by Scassellati.

5.6 Imitating head nods

By adding a tracking mechanism to the output of the face detector and then classifying these outputs, we have been able to have the system mimic yes/no head nods of the caregiver (that is, when the caretaker nods yes, the robot responds by nodding yes). The face detection module produces a stream of face locations at 20Hz. An attentional marker is attached to the most salient face stimulus, and the location of that marker is tracked from frame to frame. If the position of the marker changes drastically, or if no face is determined to be salient, then the tracking routine resets and waits for a new face to be acquired. Otherwise, the motion of the attentional marker for a fixed-duration window is classified into one of three static classes: the *yes* class, the *no* class, or the *no-motion* class. Two metrics are used to classify the motion, the cumulative sum of the displacements between frames (the relative displacement over the time window) and the cumulative sum of the absolute values of the displacements (the total distance traveled by the marker). If the horizontal total trip distance exceeds a threshold (indicating some motion), and if the horizontal cumulative displacement is below a threshold (indicating that the motion was back and forth around a mean), and if the horizontal total distance exceeds the vertical total distance, then we classify the motion as part of the *no* class. Otherwise, if the vertical cumulative total trip distance exceeds a threshold (indicating some motion), and if the vertical cumulative displacement is below a threshold (indicating that the motion was up and down around a mean), then we classify the motion as part of the *yes* class. All other motion types default to the *no-motion* class. These simple classes then drive fixed-action patterns for moving the head and eyes in a yes or no nodding motion. While this is a very simple form of imitation, it is highly selective. Merely producing horizontal or vertical movement is not sufficient for the head to mimic the action – the movement must come from a face-like object.

5.7 Regulating Interactions through Expressive Feedback

In Section 4.2, we described ongoing research toward building a robotic “infant” capable of learning communicative behaviors with the assistance of a human caretaker. For our purposes, the context for learning involves social exchanges where the robot learns how to manipulate the caretaker into satisfying the robot’s internal drives. Ultimately, the communication skills targeted for learning are those exhibited by infants such as turn taking, shared attention, and pre-linguistic vocalizations exhibiting shared meaning with the caretaker.

Towards this end, we have implemented a behavior engine for the development platform Kismet that integrates perceptions, drives, emotions, behaviors, and facial expressions. These systems influence each other to establish and maintain social interactions that can provide suitable learning episodes, i.e., where the robot is proficient yet slightly challenged, and where the robot is neither understimulated nor over-stimulated by its interaction with the human. Although we do not claim that this system parallels infants exactly, its design is heavily inspired by the role motivations and facial expressions play in maintaining an appropriate level of stimulation during social interaction with adults.

With a specific implementation, we demonstrated how the system engages in a mutually regulatory interaction with a human while distinguishing between stimuli that can be influenced socially (face stimuli) and those that cannot (motion stimuli) (Breazeal & Scassellati 1998). The total system consists of three drives (`fatigue`, `social`, and `stimulation`), three consummatory behaviors (`sleep`, `socialize`, and `play`), five emotions (`anger`, `disgust`, `fear`, `happiness`, `sadness`), two expressive states (`tiredness` and `interest`), and their corresponding facial expressions. A human interacts with the robot through direct face-to-face interaction, by waving a hand at the robot, or using a toy to play with the robot. The toys included a small plush black and white cow and an orange plastic slinky. The perceptual system classifies these interactions into two classes: *face stimuli* and *non-face stimuli*. The face detection routine classifies both the human face and the face of the plush cow as face stimuli, while the waving hand and the slinky are classified as non-face stimuli. Additionally, the motion generated by the object gives a rating of the stimulus intensity. The robot's facial expressions reflect its ongoing motivational state and provides the human with visual cues as to how to modify the interaction to keep the robot's `drives` within homeostatic ranges.

In general, as long as all the robot's drives remain within their homeostatic ranges, the robot displays `interest`. This cues the human that the interaction is of appropriate intensity. If the human engages the robot in face-to-face contact while its drives are within their homeostatic regime, the robot displays `happiness`. However, once any drive leaves its homeostatic range, the robot's `interest` and/or `happiness` wane(s) as it grows increasingly distressed. As this occurs, the robot's expression reflects its distressed state. In general, the facial expressions of the robot provide visual cues which tell whether the human should switch the type of stimulus and whether the intensity of interaction should be intensified, diminished or maintained at its current level.

For instance, if the robot is under-stimulated for an extended period of time, it shows an expression of `sadness`. This may occur either because its `social` drive has migrated into the "lonely" regime due to a lack of social stimulation (perceiving faces near by), or because its `stimulation` drive has migrated into the "bored" regime due to a lack of non-face stimulation (which could be provided by slinky motion, for instance). The expression of `sadness` upon the robot's

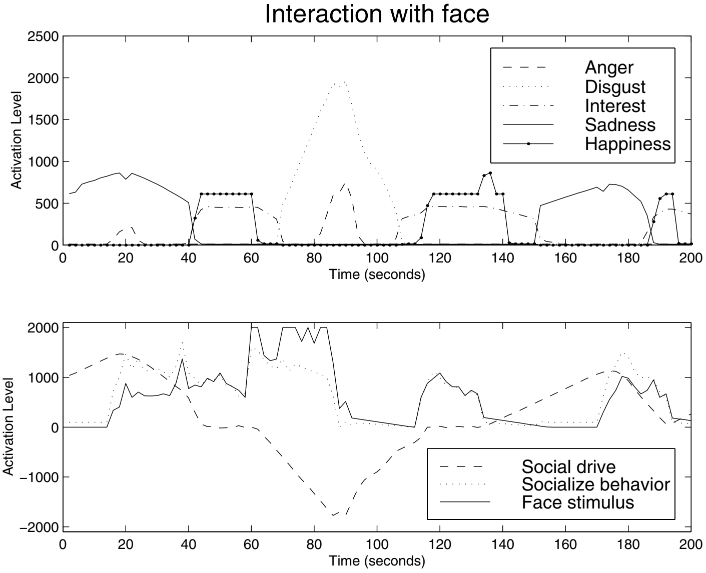

Fig. 9. Experimental results for Kismet interacting with a person's face. When the face is present and moving slowly, the robot looks interested and happy. When the face begins to move too quickly, the robot begins to show disgust, which eventually leads to anger.

face tells the caretaker that the robot needs to be played with. In contrast, if the robot receives an overly-intense face stimulus for an extended period of time, the `social` drive moves into the "asocial" regime and the robot displays an expression of `disgust`. This expression tells the caretaker that she is interacting inappropriately with the robot – moving her face too rapidly and thereby overwhelming the robot. Similarly, if the robot receives an overly-intense non-face stimulus (e.g. perceiving large slinky motions) for an extended period of time, the robot displays a look of `fear`. This expression also tells the caretaker that she is interacting inappropriately with the robot, probably moving the slinky too much and over stimulating the robot.

These interactions characterize the robot's behavior when interacting with a human. Figure 9 demonstrates how the robot's emotive cues are used to regulate the nature and intensity of social interaction, and how the nature of the interaction influences the robot's social drives and behavior. The result is an ongoing "dance" between robot and human aimed at maintaining the robot's drives within homeostatic bounds. If the robot and human are good partners, the robot remains `interested` and/or `happy` most of the time. These expressions indicate that the interaction is of appropriate intensity for learning.

6 Future Research Directions

Human beings are the most complex machines that our species has yet examined. Clearly a small effort such as that described in this paper can only scratch the surface of an understanding of how they work. We have concentrated on a number of issues that are well beyond the purely mechatronic ambitions of many robotic projects (humanoid and other). Our research has focused on exploring research issues aimed at building a fully integrated humanoid, rather than concentrating on building an integrated humanoid for its own sake.

Our ultimate goal is to understand human cognitive abilities well enough to build a humanoid robot that develops and acts similar to a person. To date, the major missing piece of our endeavor is demonstrating coherent global behavior from the existing subsystems and sub-behaviors. If all of these systems were active at once, competition for actuators and unintended couplings through the world would result in incoherence and interference among the subsystems. The problem is deeper than simply that of multi-modal systems discussed in section 4.4.

6.1 Coherence

We have used simple cues, such as visual motion and sounds, to focus the visual attention of Cog. However, each of these systems has been designed independently and assumes complete control over system resources such as actuator positions, computational resources, and sensory processing. We need to extend our current emotional and motivational models (Breazeal & Scassellati 1998) so that Cog might exhibit both a wide range of qualitatively different behaviors, and be coherent in the selection and execution of those behaviors.

It is not acceptable for Cog to be repeatedly distracted by the presence of a single person's face when trying to attend to other tasks such as grasping or manipulating an object. Looking up at a face that has just appeared in the visual field is important. Looking at what the object being manipulated is also important. Neither stimulus should completely dominate the other, but perhaps preference should be given based upon the current goals and motivations of the system. This simple example is multiplied with the square of the number of basic behaviors available to Cog, and so the problem grows rapidly. At this point neither we, nor any other robotics researchers, have focused on this problem in a way which has produced any valid solutions.

6.2 Other Perceptual Systems

We have a small number of tactile sensors mounted on Cog, but nothing near the number that occur in biological systems. Furthermore, their capabilities are quite limited when compared to the mammalian somatosensory system.

Cog does have kinesthetic sensors on some joints to provide a sense of how hard it was working, but we have not yet found a useful way to use that information. Nor have we made use of the force sensing that is available at every joint of the arms beyond direct use in feedback control — there has been no connection of that information to other cognitive mechanisms.

Finally, we have completely ignored some of the primary senses that are used by humans, especially infants; we have ignored the chemical senses of smell and taste.

Physical sensors are available for all these modalities but they are very crude compared to those that are present in humans. It may not be instructive to try to integrate these sensory modalities into Cog when the fidelity will be so much lower than that of the, admittedly crude, current modalities.

6.3 Deeper Visual Perception

So far we have managed to operate with visual capabilities that are much simpler than those of humans, although the performance of those that we do use are comparable to the best available in artificial systems. We have concentrated on motion perception, face detection and eye localization, and content-free sensory motor routines, such as smooth pursuit, the vestibular-ocular reflex, and vergence control. In addition to integrating all these pieces into a coherent whole, we must also give the system some sort of understanding of regularities in its environment.

A conventional approach to this would be to build object recognition systems and face recognition systems (as opposed to our current face detection systems). We believe that these two demands need to be addressed separately and that neither is necessarily the correct approach.

Face recognition is an obvious step beyond simple face detection. Cog should be able to invoke previous interaction patterns with particular people or toys with faces whenever that person or toy is again present in its environment. Face recognition systems typically record detailed shape or luminance information about particular faces and compare observed shape parameters against a stored database of previously seen data. We question whether moving straight to such a system is necessary and whether it might not be possible to build up a more operational sense of face recognition that may be closer to the developmental path taken by children.

In particular we suspect that rather simple measures of color and contrast patterns coupled with voice cues are sufficient to identify the handful of people and toys with which a typical infant will interact. Characteristic motion cues might also help in the recognition, leading to a stored model that is much richer than a face template for a particular person, and leading to more widespread and robust recognition of the person (or toy) from a wider range of viewpoints.

We also believe that classical object recognition techniques from machine vision are not the appropriate approach for our robot. Rather than forcing all recognition to be based on detailed shape extraction we think it is important that a developmental path for object recognition be followed. This will include development of vergence and binocularity, development of concepts of object permanence, and the early development of color perception that is robust to varied lighting.[4]

6.4 A Sense of Time

Currently, Cog has no sense of time. Everything is in the present, with the exception of some short term state implemented via the emotional levels present in the Kismet platform. These emotional states can act as the keys to K-line like indexing into associative memory, but this is not sufficient to produce the richness of experience and subsequent intelligence that humans exhibit.

A key technical problem is how to relate the essentially static and timeless aspects of memory that are present in neural networks, registration maps, self-organizing maps, nearest neighbor approximations, and associative memory, to the flow of time we as human beings experience.

This is a real technical challenge. A conventional AI system has separate program and data, and the program has a natural flow of time that it can then record in a data structure. Our models do not make this sort of distinction; there is neither a sequential place in memory nor a process to capitalize on it. Given that we have rejected the conventional approaches, we must find a solution to the problem of how episodic memory might arise.

This chapter has focused on the current capabilities of our humanoid robotic systems and the future directions that our research will address. These problems are simply the beginning of what we hope will be a rich source of both new research questions and innovative solutions to existing problems.

7 Acknowledgments

Support for this project is provided in part by an ONR/ARPA Vision MURI Grant (No. N00014-95-1-0600).

References

An, C. H., Atkeson, C. G. & Hollerbach, J. M. (1988), *Model-based control of a robot manipulator*, MIT Press, Cambridge, MA.

Ashby, W. R. (1960), *Design for a Brain*, second edn, Chapman and Hall.

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

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

Blythe, J. & Veloso, M. (1997), Analogical Replay for Efficient Conditional Planning, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 668–673.

[4] It is well known that the human visual system, at least in adults, is sensitive to the actual pigmentation of surfaces rather than the frequency spectrum of the light that arrives on the retina. This is a remarkable and counter-intuitive fact, and is rarely used in modern computer vision, where cheap successes with simple direct color segmentation have gotten impressive but non-extensible results.

Boutilier, C. & Brafman, R. I. (1997), Planning with Concurrent Interacting Actions, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 720–726.

Brafman, R. I. (1997), A Heuristic Variable Grid Solution Method for POMDPs, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 727–733.

Breazeal, C. & Scassellati, B. (1998), 'Infant-like Social Interactions between a Robot and a Human Caretaker', *Adaptive Behavior*. In submission.

Breazeal, C. & Velasquez, J. (1998), Toward teaching a robot "infant" using emotive communication acts, *in* 'Socially Situated Intelligence: Papers from the 1998 Simulated Adaptive Behavior Workshop'.

Breazeal (Ferrell), C. (1998), A Motivational System for Regulating Human-Robot Interaction, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

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

Bullowa, M. (1979), *Before Speech: The Beginning of Interpersonal Communicaion*, Cambridge University Press, Cambridge, London.

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

Chappell, P. & Sander, L. (1979), Mutual regulation of the neonatal-materal interactive process: context for the origins of communication, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 191–206.

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

Costello, T. (1997), Beyond Minimizing Change, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 448–453.

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

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.

Ferrell, C. (1996), Orientation Behavior using Registered Topographic Maps, *in* 'From Animals to Animats: Proceedings of 1996 Society of Adaptive Behavior', Cape Cod, Massachusetts, pp. 94–103.

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.

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

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

Ghez, C. (1992), Posture, *in* E. R. Kandel, J. H. Schwartz & T. M. Jessell, eds, 'Principles of Neural Science', 3rd edn, Appleton and Lange.

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.

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

Halliday, M. (1975), *Learning How to Mean: Explorations in the Development of Language*, Elsevier, New York, NY.

Hatsopoulos, N. G. & Warren, W. H. (1996), 'Resonance Tuning in Rhythmic Arm Movements', *Journal of Motor Behavior* **28**(1), 3–14.

Hauskrecht, M. (1997), Incremental Methods for computing bounds in partially observable Markov decision processes, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 734–739.

Herr, H. (1993), Human Powered Elastic Mechanisms, Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Hirai, K., Hirose, M., Haikawa, Y. & Takenaka, T. (1998), The Development of the Honda Humanoid Robot, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

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

Irie, R. E. (1997), Multimodal Sensory Integration for Localization in a Humanoid Robot, *in* 'Proceedings of Second IJCAI Workshop on Computational Auditory Scene Analysis (CASA'97)', IJCAI-97.

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

Kanehiro, F., Mizuuchi, I., Koyasako, K., Kakiuchi, Y., Inaba, M. & Inoue, H. (1998), Development of a Remote-Brained Humanoid for Research on Whole Body Action, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

Kaye, K. (1979), Thickening Thin Data: The Maternal Role in Developing Communication and Language, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 191–206.

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

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

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

Littman, M. L. (1997), Probabilistic Propositional Planning: Representations and Complexity, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 748–754.

Lobo, J., Mendez, G. & Taylor, S. R. (1997), Adding Knowledge to the Action Description Language $\mathcal{A}$, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 454–459.

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. 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', Cape Cod, Massachusetts, pp. 35–44.

Mason, M. T. & Salisbury, Jr., J. K. (1985), *Robot Hands and the Mechanics of Manipulation*, MIT Press, Cambridge, Massachusetts.

Matsuoka, K. (1985), 'Sustained oscillations generated by mutually inhibiting neurons with adaption', *Biological Cybernetics* **52**, 367–376.

Matsuoka, K. (1987), 'Mechanisms of frequency and pattern control in neural rhythm generators', *Biological Cybernetics* **56**, 345–353.

McCain, N. & Turner, H. (1997), Causal Theories of Action and Change, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-97)', pp. 460–465.

McGeer, T. (1990), Passive Walking with Knees, *in* 'Proc 1990 IEEE Intl Conf on Robotics and Automation'.

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

Morita, T., Shibuya, K. & Sugano, S. (1998), Design and Control of Mobile Manipulation System for Human Symbiotic Humanoid, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

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.

Newson, J. (1979), The growth of shared understandings between infant and caregiver, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 207–222.

Panerai, F. & Sandini, G. (1998), 'Oculo-Motor Stabilization Reflexes: Integration of Inertial and Visual Information', *Neural Networks.* In press.

Peskin, J. & Scassellati, B. (1997), Image Stabilization through Vestibular and Retinal Feedback, *in* R. Brooks, ed., 'Research Abstracts', MIT Artificial Intelligence Laboratory.

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.

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.

Rosenbaum, D. A. et al. (1993), 'Knowledge Model for Selecting and Producing Reaching Movements', *Journal of Motor Behavior* **25**(3), 217–227.

Salisbury, J., Townsend, W. T., Eberman, B. S. & DiPietro, D. M. (1988), Preliminary Design of a Whole arm Manipulation System (WAMS), *in* 'Proc 1988 IEEE Intl Conf on Robotics and Automation'.

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

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

Scassellati, B. (1998*b*), Building Behaviors Developmentally: A New Formalism, *in* 'Integrating Robotics Research: Papers from the 1998 AAAI Spring Symposium', AAAI Press.

Scassellati, B. (1998*c*), Finding Eyes and Faces with a Foveated Vision System, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

Scassellati, B. (1998*d*), Imitation and Mechanisms of Shared Attention: A Developmental Structure for Building Social Skills, *in* 'Agents in Interaction - Acquiring Competence through Imitation: Papers from a Workshop at the Second International Conference on Autonomous Agents'.

Schaal, S. & Atkeson, C. G. (1993), Open loop Stable Control Strategies for Robot Juggling, *in* 'Proceedings 1993 IEEE International Conference on Robotics and Automation', Vol. 3, pp. 913–918.

Schneider, K., Zernicke, R. F., Schmidt, R. A. & Hart, T. J. (1989), 'Changes in limb dynamics during the practice of rapid arm movements', *Journal of Biomechanics* **22**(8–9), 805–817.

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

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

Takanishi, A., Hirano, S. & Sato, K. (1998), Development of an anthropomorhpic Head-Eye System for a Humanoid Robot, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

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

Trevarthen, C. (1979), Communication and cooperation in early infancy: a description of primary intersubjectivity, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 321–348.

Tronick, E., Als, H. & Adamson, L. (1979), Structure of early Face-to-Face Communicative Interactions, *in* M. Bullowa, ed., 'Before Speech', Cambridge University Press, pp. 349–370.

Warren, C. A. & Karrer, R. (1984), 'Movement-related potentials during development: A replication and extension of relationships to age, motor control, mental status and IQ', *International Journal of Neuroscience* **1984**, 81–96.

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

Weiskrantz, L. (1986), *Blindsight: A Case Study and Implications*, Clarendon Press, Oxford.

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

Williamson, M. M. (1996), Postural Primitives: Interactive Behavior for a Humanoid Robot Arm, *in* 'Fourth International Conference on Simulation of Adaptive Behavior', Cape Cod, Massachusetts, pp. 124–131.

Williamson, M. M. (1998*a*), Exploiting natural dynamics in robot control, *in* 'Fourteenth European Meeting on Cybernetics and Systems Research (EMCSR '98)', Vienna, Austria.

Williamson, M. M. (1998*b*), Rhythmic robot control using oscillators, *in* 'IROS '98'. Submitted.

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

Yamato, J. (1998), Tracking moving object by stereo vision head with vergence for humanoid robot, Master's thesis, MIT.

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.