Technologies for Human/Humanoid Natural Interactions

Rodney A. Brooks, Cynthia Breazeal, Brian Scassellati, Una-May O'Reilly
MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139, USA
{brooks, cynthia, scaz, unamay}@ai.mit.edu

Abstract

There are a number of reasons to be interested in building humanoid robots. They include (1) since almost all human artifacts have been designed to easy for humans to interact with, humanoid robots provide backward compatibility with the existing human constructed world, (2) humanoid robots provide a natural form for humans to operate through telepresence since they have the same kinematic design as humans themselves, (3) by building humanoid robots that model humans directly they will be a useful tool in understanding how humans develop and operate as they provide a platform for experimenting with different hypotheses about humans and (4) humanoid robots, given sufficient abilities, will present a natural interface to people and people will be able to use their instinctive and culturally developed subconscious techniques for communicating with other people to communicate with humanoid robots. In this paper we take reason (4) seriously, and examine some of the technologies that are necessary to make this hope a reality.

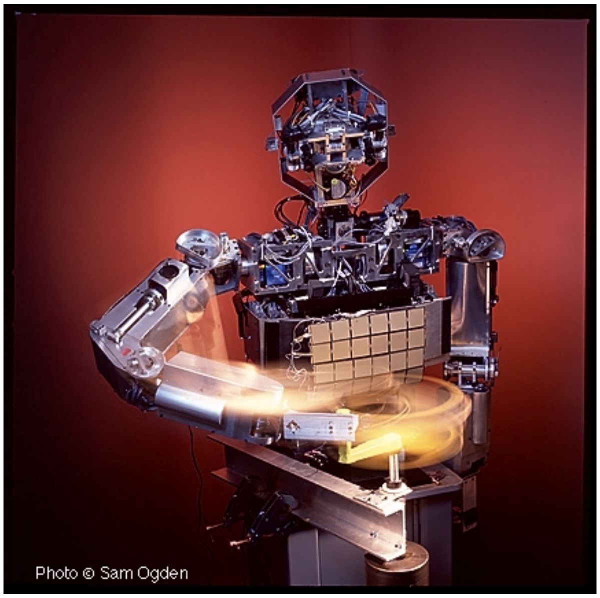

Figure 1: The robot Cog, using neural oscillators to carry out a complex manipulation task without any model of kinematics or dynamics.

1 Scenarios

We outline three scenarios where a humanoid robot (Brooks 1996) might be used in the future. In each case the robot is to be instructed by a person on how to carry out its tasks. By using natural communications methods the instruction process is easy for the person, and does not require that they be specially trained to instruct a robot. The highlighted numbers refer to the numbered subsections of section 2 following these three scenarios. These subsections illuminate the research issues that must be solved in order to build humanoid robots that can perform as suggested in the scenarios. We also give a summary of our own results in these directions using our robots Cog and Kismet, pictured in figures 1 and 2.

[a] Opening a gas tank

Suppose a robot is to be tasked with filling the gas tanks at a gas station, where many different makes and models of cars, SUVs, and trucks will need to be serviced. The human instructor calls the robot over to an example automobile, and instructs it to watch how to open the gas tank. The person looks over his shoulder to confirm that the robot is looking in the right place *(1–the robot should not imitate this and look over its shoulder every time it opens a gas tank)*. The person flips open the gas filler cover door with one hand and glances up at the robot *(6–the robot must give some sign, through nodding, or vocalizing, that it is following)*, then twists off the gas cap with his other hand *(2–the robot needs to realize that it*

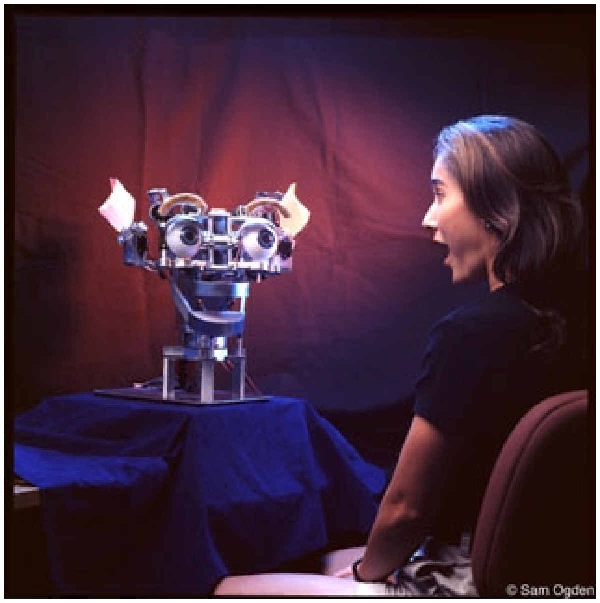

Figure 2: The robot Kismet, interacting with a person through facial expressions.

can use its one and only arm to do both tasks, 4–the robot must chain together the two sequential actions, but then later not be confused as to what to do first on a vehicle that is missing the cover door, 1,2–the robot notes the way the person's elbow swings about as he is twisting the cap, but must realize that it need not swing its own elbow to achieve the same result as the kinematics of its arm are different from those of a human). The person hands the tank cap to the robot and tells it to put things back *(4–now the robot must reverse the chain of events, 5–generalize the individual steps so that it can reverse them, and 3–evaluate how well it does the job and improve its performance over time; the robot must also generalize the task to other models of cars when it is confronted with them)*.

[b] Earthquake reconnaissance

In an earthquake disaster a reconnaissance robot is to be sent down a street between unsafe tall buildings and off to the right at a particular intersection to check for signs of survivors in a side street. There is heaving construction equipment moving concrete slabs, and the situation is noisy and confusing. The operator gestures down the street and yells to the robot to go that way *(6–the robot must give some indication that it is following what is being said, especially given the noisy situation, 2–the robot must understand the command as a gesture telling it what to do with its body as a whole, not asking it to point its hand in the same direction)*, then waves his hand to the right *(6–the robot acknowledges)*, and tells the robot to turn at the intersection with the crushed red car partially blocking it *(6–the robot gestures to the right, looking for confirmation from the person on the required direction, and the operator instinctively nods his head "yes")*, and to check for survivors and report back *(4–the robot must understand the sequence of operations that it is to do, and when it is in the side street and the intersection has a broken water main making it impossible to return by the route it came, 5–the robot must be able to generalize that the goal is to get back to the operator, and 3–re-plan a different homeward path)*.

[c] Loading a truck

A humanoid robot is to drive a fork-lift vehicle to load a truck with un-palletized bags of rice during a humanitarian relief operation. The human shows the robot the pile of bags, then takes the robot over to the truck *(1–the robot realizes that this is all part of the instruction process)*, then uses hand gestures to indicate the desired orientation of the bags *(2–so the robot must be able to map the hand motions to the relative dimensions of the sides of the bag, and 1–understand that these placements are the principal things it must imitate in loading the truck)*, and uses placing motions to indicate how they should be layed down in a regular pattern *(5–the robot must generalize from a few places to a pattern which covers the whole truck floor, and 4–it must chain the placement actions and the transport actions into a more complex sequence of actions)*, and all the time glances at the robot to receive confirmation that it is understanding the instructions *(6– the robot must interpret these glances as the cues for when it should give either positive or negative feedback to the person that it is understanding what he is saying)*.

2 Issues

Each of the scenarios previously described raises a number of difficult practical and research issues. We believe the following six research areas are the critical ones.

2.1 Knowing what to imitate

One of the most difficult problems in complex robotic systems is determining which of the incoming sensory signals are relevant to the current task. The robot must both segment the incoming visual signals into salient objects and determine which are relevant to the task at hand. For example, in scenario [**a**], the robot must observe the entire scene and segment it into salient objects (such as the instructor's hand, the cover to the gas tank, the vehicle's tires, a side of the vehicle) and actions (the instructor's moving hand twisting the gas cap, the movement of a tree blowing in the wind in the background, and the instructor's head turning toward the robot). The robot must determine which of these objects and events are necessary to the task at hand (such as the gas tank and the movement of the instructor's elbow), which events and actions are important to the instructional process but not to the task itself (such as the movement of the instructor's head), and which are inconsequential (such as the instructor wiping his brow or the movement of the trees in the background). The robot must also determine to what extent each action must be imitated. For example, in scenario [**c**], the robot must determine that a relevant part of the task is to load bags but that the exact ways in which the instructor handles the bags or the instructor's posture while lifting need not be mimicked.

We have been developing techniques to solve these problems: recognizing inherent saliency in objects and recognizing objects and actions that are salient because the instructor is attending to them. Measures of inherent object saliency, (e.g. color, texture, and face detection) can be combined with an attentional system to generate initial estimations of relevant objects. Additional refinement can be obtained by observing the attentional states and social actions of the instructor.

2.2 Mapping between bodies

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

To constrain the space of potential mappings we will use the connection between how events are sensed and the reactions they generate. By also attending to both the static properties of objects and the current social situation, the number of potential motor responses can be limited. For example, in scenario [**c**], the size and shape of the bag limits the number of ways in which the robot can handle the bag while the social cues from the instructor constrains the potential responses.

2.3 Recognizing success

Once a robot can observe an action and attempt to imitate it, how can the robot determine whether or not it has been successful? Further, if the robot has been unsuccessful, how does it determine which parts of its performance were inadequate? If the robot is attempting to load bags into a truck as in scenario [**c**], has the robot been successful if it picks up the bag that the instructor has already loaded and moves it to a different position in the truck? Is the robot successful if it picks up a new bag, attempts to place the bag in the same space as the bag that the instructor already loaded and in doing so pushes the other bag out of the truck? Is the robot successful if it picks up a new bag and loads it on top of another bag in an unstable pile? In all of these cases, how does the robot determine which parts of its actions have been inadequate?

In the case of imitation, the difficulty of obtaining a success criterion can be simplified by exploiting the natural structure of social interactions. As the robot performs its task, the facial expressions, vocalizations, and actions of the instructor all provide feedback that will allow the robot to determine whether or not it has achieved the desired goal. Imitation is also an iterative process; the instructor demonstrates, the student performs, and then the instructor demonstrates again, often exaggerating or focusing on aspects of the task that were not performed successfully. By repeatedly responding to the social cues that initially allowed the robot to understand and identify which salient aspects of the scene to imitate, the robot can incrementally refine its approximation of the actions of the instructor.

2.4 Chaining

To perform many goal-oriented tasks a robot must be capable of chaining together imitations of simpler tasks. Combining simple tasks into a flexible and robust action plan is vital for success. For example, in scenario [**b**], the robot must maintain a sequence of directions (going down the street, turning at the intersection, checking for survivors, and returning to report) which must be performed in order, and each of which consists of a number of sub-parts that must be performed in sequence. Each of these behaviors must be flexible to deal with unforeseen situations while remaining robust to the changes in the environment. The behaviors must also be combined in a flexible manner; depending on environmental conditions, the order of the behaviors may need to be altered or certain behaviors may need to be omitted. For example, if the intersection is blocked, a new route must be found.

Recognizing what sequence of actions is necessary can be simplified by recognizing the social context of the situation. In the same way that social cues limit motor actions, social signals indicate which actions can be performed at a particular time. Low-level behavior driven responses combined with high-level planned actions provide robustness and flexibility.

2.5 Generalizing

Once a robot has learned to imitate an action, it should be able to utilize that skill to simplify learning other tasks. For example, in scenario [**c**], once the robot has learned to load bags into the truck, learning to load boxes should be simplified by the previously acquired knowledge. The robot must have the ability to recognize and modify applicable skills that it has acquired. It should also be capable of extracting invariants about the world from its interactions and thus acquire "common-sense" knowledge. For example, the robot might apply the concept of support from learning to stack boxes in a truck to assembly tasks.

The robot must be capable of taking learned sequences and treating them as "primitives" for combinations into even more complex behaviors while still allowing the feedback from the instructor to drive optimization of individual sub-parts. For example, in [**c**] the robot could learn that a sequence for picking up an object has the same overall structure applicable to different types of objects (all involve approaching an object, moving your arms to a point near the object and then grasping it) but might require different optimizations to the sub-components (picking up a bag of rice requires a different grasp than picking up a box of provisions).

2.6 Making interactions intuitive

To make a robotic system useful, it must have a simple and intuitive interface. Our goal is to build robotic systems that can capitalize on the natural social signals humans subconsciously use in communication with each other. For example, in [**a**], the robot must provide feedback to the instructor that it has understood its instructions, signal when it does not understand, extrapolate the importance of each piece of the task based on the instructor's emotional cues, and recognize the start and end of the instruction period. A system that operates using natural human social expressions allows anyone, without prior instruction, to instruct the robot in a simple, natural, and intuitive manner. Because so many of these social signals are completely unconscious for humans, the task of teaching is simplified.

Utilizing these signals requires highly specialized perceptual systems that are sensitive to the types of social cues that humans use. For example, humans are extremely sensitive to the direction of gaze, the tone of voice, and facial expressions. Recognizing these cues in natural environments requires high accuracy (such as finding the angle of gaze) and high speed (so that it can respond quickly enough to maintain social convention) without relying upon simplified environmental invariants.

3 Coordinating the issues

We have attacked these six issues by carrying out work under four research themes.

I We have built systems that allow robots to engage in *social interactions* with humans by utilizing normal cross-cultural human-to-human sublinguistic interactions.

II We organize the subsystems of our robots to follow human-like *developmental paths* so that we can exploit the solutions that evolution has painstakingly discovered. A developmental strategy allows increasingly more complex skills and competencies to be layered on top of continuously self-calibrating simpler competencies. It also organizes knowledge of the world in the same manner as that used by humans and thus contributes to intuitive understanding.

III Our subsystems exploit the *embodiment* in the world, of the robots they control. Embodiment facilitates a robot using the world as a tool for organizing and manipulating knowledge. At the same time, just as with humans, it directs choices away from what could be done only with great difficulty and towards what can be done naturally and in keeping with the physics of the world.

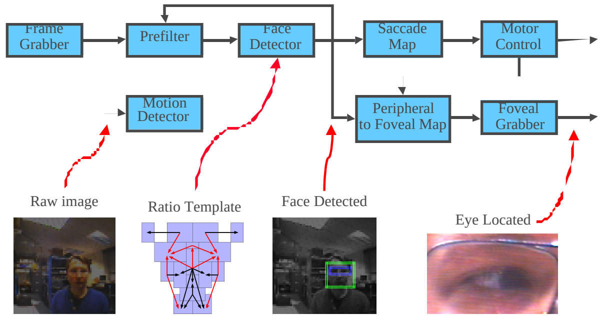

Figure 3: The first stage of gaze detection is finding faces and eyes. This shows the system we have built for both Cog and Kismet which finds faces in the periphery, saccades to them, them and knows where to look for the eyes.

IV Our humanoid robots are highly *integrated* which maximize the efficiency and accuracy of complementary sensory and motor systems.

In the following subsections we outline these four methodologies and explicitly relate their techniques to the solutions of our six key problems.

3.I Social interaction

Humans socially engage each other for assistance, teaching, and information. People are able to communicate rapidly, efficiently, and fluidly because we share the same set of social conventions, display and interpret common social cues, and possess commonsense social knowledge. Furthermore, humans instinctively instruct and learn from each other in social scenarios. New skills are transferred between people through mimicry or imitation, through direct tutelage, or by scaffolding (actively assisting the learning process by reducing distractions, directing attention to the task's critical attributes, decomposing the task into more manageable chunks, or providing rich and ongoing reinforcement).

Similarly, social interaction can be used as a means of natural human-machine interaction and a powerful way for transferring important skills, tasks, and information to a robot. Were a robot to possess similar social skills and knowledge as people, humans would have a natural way of communicating effectively with the robot and vice versa. Tasking the robot would not require any special training, and would be as natural as tasking any other person. Furthermore, a socially competent robot could take advantage of the same sorts of social learning and teaching scenarios that humans readily engage. Hence, people could instruct the robot in an intuitive manner and the robot could take advantage of various social cues to facilitate its own learning of a task.

Facilitates knowing what to imitate. Social interaction plays a critical role in facilitating imitative forms of learning. Fundamental social cues, such as gaze direction (see figure 3), and basic social skills, such as shared attention, can be used by a robot to determine the important features of the task. Human instructors naturally attend to the key aspects of a task when demonstrating a task to someone else. For example, in [**a**], the person will naturally look at the gas filler cover door as he flips it open and will watch his own hand as he twists off the cap. By directing its own attention to the object of the human's attention the robot will automatically attend to the critical aspects of the task. The robot's gaze direction can also serve as an important feedback signal for the human; the person looks over his shoulder to confirm that the robot is looking in the right place. If this is not the case, then the person can actively direct the robot's attention to the gas tank cover, perhaps by pointing to it or tapping on it. In general, if the robot has basic social knowledge, then it will be able to distinguish acts for communication from acts directly related to the task being taught.

Facilitates knowing when you have it right. Social interaction can also play a critical role in helping the robot identify the relevant success criteria for a task as well as identifying when success has been achieved. Human instructors serve as natural progress estimators and progress evaluators to a person learning a task. Typically this information is given through facial expressions (smiles or frowns), gestures (nodding or shaking of the head) and verbal feedback ("Yes, that's right.", "No, not quite.").

Without human instruction, designing suitable reinforcement functions or progress estimators for robots is a notoriously difficult problem that often leads to learning brittle behaviours. This aspect of the learning problem could be greatly facilitated if the robot could exploit the instructor's social feedback cues, query the instructor or make use of readily available feedback. Humans naturally query their instructor by simply glancing back to his face with an inquisitive expression. The robot could use the same social skill to query the human instructor, as illustrated in scenario [**b**].

Facilitates the teaching process. In general, a wide variety of social cues can play a powerful role in tuning the teaching process to be suitable for the

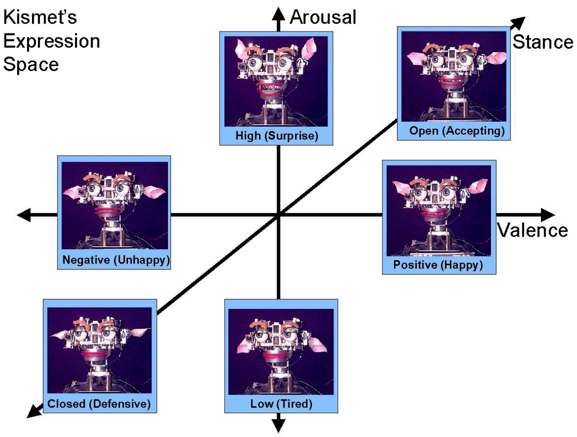

Figure 4: Kismet is able to display a wide range of facial expressions, based on an underlying three dimensional expression space.

learner. Basic social skills such as turn taking are critical for social learning; instruction is often an iterative process where the instructor and student alternate turns. The instructor continually modifies the way he performs the task, perhaps exaggerating those aspects that the student performed inadequately, in an effort to refine the student's subsequent performance.

Furthermore, while taking his turn, the instructor often looks to the student's face to determine whether the student appears confused or understands what is being demonstrated. Expressive displays are effortlessly used by people to control the rate of information exchange, to speed it up, slow it down, or elaborate as appropriate. If the student appears confused, the instructor slows down the training scenario until the student is ready to proceed. Hence, displaying facial expressions (figure 4) is an important cue for the instructor as well as the student. Overall, the ability to take turns and display expressive feedback are important social skills for a robot to possess if it is to participate in this sort of natural training process. This is illustrated in scenario **[c]** where the human looks to the robot for confirmation that it understands what it is being shown. Without these skills, it will be difficult for the human instructor to promote and maintain an appropriate learning environment for the robot. This may result in wasting time teaching the robot what it already understands, or proceeding at such a fast rate that the robot is unable to learn the task.

Promotes an intuitive interface. Above, we have argued how social interaction can facilitate both the learning process for the robot as well as the teaching process for the human. The argument can be easily extended to include building social skills and knowledge into a robot so that it is easy to communicate with in general. This includes having the robot more effectively understand the human as well as having the human more effectively understand the robot. By implementing cross-cultural sub-linguistic interactions, the robot will more readily be able to express its internal state (emotions, drives, goals) in a manner that is intuitively understood by people without relying on some artificial vocabulary. A robot with social knowledge can recognize the goals and desires of others and more accurately react to the emotional, attentional, and cognitive states of the observer, learn to anticipate the reactions of the observer, and modify its own behavior accordingly. For example, in scenario **[b]**, the robot must comprehend the urgency and risk of its mission, as communicated by its instructor, in order to direct its choice of task execution.

Current progress. A machine that can interact socially with a human instructor requires a variety of perceptual, cognitive, and motor skills. Over the last several years our research group has begun building the foundational skills for three aspects of social interaction: recognizing and responding to attentional cues, producing and recognizing emotional expression, and regulating interactions to facilitate learning. Interaction intensity is regulated through displayed facial expressions. The robot's reactions to external stimuli depend upon its internal motivational state, the quality of the incoming stimulus and the history of interaction.

Recognizing the attentional states of the instructor assists in knowing what to imitate *(1)*, evaluating success *(3)*, and in providing an intuitive interface *(6)*. To allow our robots to recognize attentional states, we have already implemented a face and eye detection system (see figure 3), which allows the robot to detect faces, move its eyes to foveate the face, and extract a high-resolution image of the eye (Scassellati 1998). We are currently adding image analysis and geometric interpolation algorithms to extract the direction of gaze and interpolate to an object in the world (Scassellati 1999). We propose extending this system to recognize other forms of joint attention (such as pointing) and to produce predictive models of the instructor's goals, beliefs, and desires.

Identifying the emotional states of the instructor and responding with its own emotional displays will assist our robot in knowing what to imitate *(1)*, evaluating success *(3)*, and in providing a natural interface *(6)*. We have developed robots with the ability to display facial expressions and have developed emotional models that drive them based upon environmental stimuli and internal motivations (Breazeal & Scassellati 2000, Breazeal & Scassellati 1999). We have been exploring the role of emotions in learning, and have successfully demonstrated fear conditioning (using color and sound stimuli) on a mobile robot (Velasquez 1998). We are continuing to develop these ideas to eventually train a robot using emotive feedback from the instructor (Breazeal & Velasquez 1998).

To successfully learn any social task, the robot must also be capable of regulating the rate and intensity of instruction to match its current understanding and capabilities *(3)*. The robot Kismet differentiates between salient social stimuli (things that have faces) and those that are interesting but non-social (such as brightly colored or moving objects). Just as infants manipulate their parents, Kismet can utilize its facial expressions to naturally influence the rate and content of the instructor's lessons. For example, if the instructor is moving too quickly, the robot responds with a frustrated and angry expression. These social cues are unconsciously interpreted by the instructor, who modifies his behavior to maintain the interaction. We propose extending this capability to allow the robot to optimize its own learning environment through social manipulations of the instructor.

We have also developed an attentional system that integrates motion, color, and face saliency cues (Breazeal & Scassellati 1999). A saliency map is computed for each feature, and these are combined by a weighted average into an attentional map. The attentional map represents a landscape, where regions of high saliency are represened as peaks in this map. A habituation mechanisms is included so that the robot is not held indefinitly captive on a particular stimulus. Motivational factors (drives, emotions, behaviors) can influence the gains, to heighten attention to those features that are particularly behaviorally relevent to the robot at that time. The robot's eyes fixate on the most salient stimuli.

3.II Developmental approach

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. In a similar way, we do not expect our robots to perform completely correctly without any experience in the world. We have been studying human development both as a tool for building robotic systems and as a technique which facilitates learning.

The most important contribution of a developmental methodology is that examples of structured skill decomposition and shows how the complexity of a task can be gradually increased in step with the competency of the system. Human development provides us with insight into how complex behaviors and skills (such as manipulating an object such as a gas cap in [**a**] or perceiving that the instructor's attention is focused on a particular bag of food in [**c**]) can be broken down into simpler behaviors. Already acquired sub-skills and knowledge are re-usable, place simplifying constraints on ongoing skill acquisition, and minimize the quantity of new information that must be acquired. By exploiting a gradual increase in both internal complexity (perceptual and motor) and external complexity (task and environmental complexity regulated by the instructor), while reusing structures and information gained from previously learned behaviors, we hope to be able to learn increasingly sophisticated behaviors.

Development simplifies knowing what to imitate. A developmental approach keeps the necessary perceptual tasks in step with gradually increasing capabilities and optimizes learning by matching the complexity of the task with the current capabilities of the system. For example, infants are born with limited visual input (low acuity). Their visual performance develops in step with their ability to process the influx of stimulation (Johnson 1993). By having limited quality and types of perceptual information, infants are forced first to learn skills loosely and then to refine those skills as they develop better perception. In a similar way, our robotic systems will first utilize simpler perceptual abilities to recognize the general perceptual qualities (such as object position and motion) which will gradually be refined with more complex perceptual properties (such as better resolution vision, more complex auditory scene analysis, face detection, etc.). This allows us to first concentrate on imitating the overall scene properties such as moving a bag of food from one place to another in [**c**] without getting lost in the details of the action.

Development facilitates mapping between bodies. A developmental approach simplifies the mapping problem *(2)* by providing a methodology for incremental refinement of the perceptual and motor mapping. In human development, 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). Using a method of incremental refinement, our robots will first learn to imitate large-scale motions indicated by the instructor. Once the motion has been successfully learned and optimized at this rough granularity, the robot will begin to refine its approximation of the movement at a finer granularity. For example, the action of opening a gas tank cover in scenario **[a]** can first be learned by attending only to the gross movements of the instructor's arm and later refined to match detailed movements of the wrist and elbow.

Development provides a "road-map" for chaining and generalizing behaviors. A developmental methodology also provides a "road-map" of how simple skills combine to build complex skills. One problem that is encountered both with chaining together actions *(4)* and with generalizing imitated actions into different and more complex tasks *(5)* is recognizing the appropriate decomposition of the complex action. For example, in learning to reach out and grasp an object (such as the bag of food in scenario **[c]**, or the gas tank cover in scenario **[a]**), we must identify the smaller action components that make up this intermediate-level task. By studying the development of reaching and grasping in human infants we have obtained not only a set of potential behavior primitives but also one way to combine these primitives into more complex behaviors (Diamond 1990). By examining the ways that evolution has combined skills into complex behaviors, we gain valuable insight on ways to decompose our robotic problems.

Development promotes an intuitive interface. Building a system that can recognize and produce complex social behaviors such as cross-cultural cues *(6)* requires a skill decomposition that maintains the complexity and richness of the behaviors while still being simple enough to implement. Evidence from the development of these non-verbal social skills in children (Hobson 1993) and autistics (Baron-Cohen 1995, Frith 1990), and evolutionary studies of non-verbal communication (Povinelli & Preuss 1995), all demonstrate that the development of complex social skills can be decomposed into a sequence of simpler behaviors. The basis of this developmental chain is a set of behaviors that allow an individual to share with another person the experience of a third object (Wood, Bruner & Ross 1976). For example, the student might point to an object, or alternate between looking at the instructor and the object. These techniques for obtaining joint (or shared) attention follow a strict developmental progression beginning with detection of eye contact, incremental refinement of eye gaze detection, recognition of pointing gestures, and culminating in the ability to attribute beliefs, desires, and goals to other individuals.

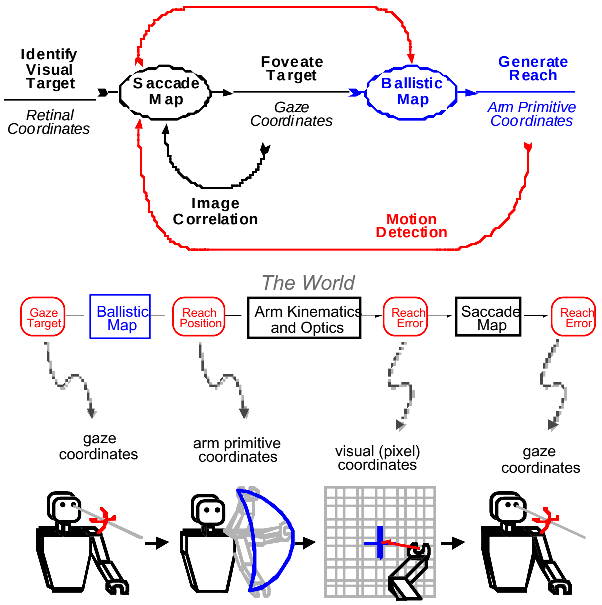

Figure 5: A developmental approach allows efficient re-use of knowledge. Learning how to saccade to a target facilitates learning to reach toward a target. The saccade map is used both to get the eyes to the target, and to interpret where the arm ends up in terms of where the robot would have had to be looking for that to be a good reach. This provides an error signal in the right coordinate system (gaze coordinates) in order to learn the reaching map.

Current progress. Developmental methodologies provide us with decomposition methods and incremental strategies for refining robotic perceptual, motor, and cognitive abilities. Building a robotic system that can imitate the actions of an instructor requires basic eye motor skills, face and gaze detection, determination of eye direction, gesture recognition, attentional systems that allow for social behavior selection at appropriate moments, emotive responses, arm motor control, gaze stabilization, and many other skills.

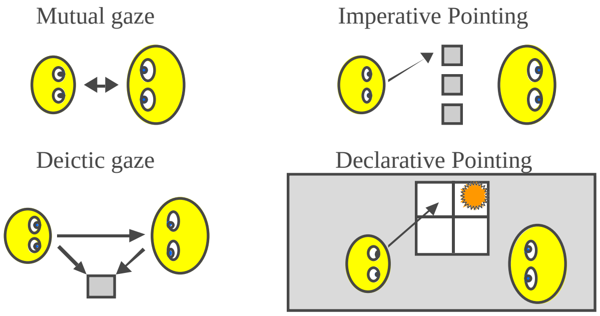

Figure 6: Stages of development of joint attention skills.

Figure 7: Cog imitating head nods.

These perceptual, motor, and social skills can all be staged within a developmental framework which decomposes our problems into manageable pieces and promises the benefits of incremental learning.

We have already begun studying motor systems that follow developmental paths similar to those in humans. Marjanović, Scassellati & Williamson (1996) applied a developmental technique to the problem of reaching for a visual object, a precursor to manipulative tasks. Following the developmental path that Diamond (1990) demonstrated in infants between five and twelve months of age, Marjanović et al. (1996) implemented a pointing behavior for the humanoid robot Cog. The robot first detects moving objects (a simple saliency metric), foveates the object, and then reaches for the object with its six degree-of-freedom arm. The robot learns this behavior incrementally over a period of a few hours, using gradient descent methods to train forward and inverse mappings between a visual parameter space and an arm position parameter space without human supervision. The learning is done in stages, first isolating the foveation behavior and then adding additional degrees of freedom as performance improves. This developmental progression simplified the complex reaching behavior by leveraging knowledge that was acquired in earlier stages (specifically, the foveation behavior) to provide error signals that allow learning a forward and inverse kinematic mapping as illustrated in figure 5. We will extend this work to allow multi-arm manipulation, more complex forms of reaching (such as reaching around an object), and more robust grasping techniques.

We have also already begun to address the problems of learning social skills in a developmental setting. Scassellati (1996) has discussed how a humanoid robot might acquire basic social competencies through this sort of developmental methodology (figure 6). To enable our robot to recognize and maintain eye contact, we have implemented a perceptual system capable of finding faces and eyes (Scassellati 1998). 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 is identified, the robot saccades to that target using a learned visual-motor mapping. The location of the face in peripheral image coordinates is mapped into foveal image coordinates using a second learned mapping. The location of the face within the foveal image is used to extract the sub-image containing the eye.

In building the basic social skills of joint attention, we have also identified an unexpected benefit of the developmental methodology: the availability of closely related skills. For example, simply 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 instructor (that is, when the instructor nods yes, the robot responds by nodding yes; see figure ??). The robot classifies the output of the face detector and responds with a fixed-action pattern 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. Because our developmental methodology requires us to construct many sub-skills that are useful in a variety of environmental situations, we believe that these primitive behaviors and skills can be utilized in a variety of circumstances.

3.III Embodiment

The distinctive robots that are to be used in this research each have human-like aspects of their bodies. First, each robot is equipped with a human-like robotic head and neck with sensing elements that are analogous to human eyes and ears. Second, each robot has a face with mechanisms that allow for active control of expressive facial features. Third, Cog's body is similar to a human's entire upper-body including human-like arms. Having a human-like shape and dynamics confers advantages which are described in the following section.

Embodiment helps you know what to imitate. In standard instruction for a physical task, the student needs to use essentially the same solution as the instructor. This constrains the space of possible solutions to ones similar to the instructor's solution. Having a similar physical body thus makes deciding what to imitate an easier task *(1)*. For example, when opening the gas cap, the instructor gives an initial configuration for the arm. If the robot has the same shape as the human, it can copy this configuration as a starting point for its solution. Conversely, a different morphology would imply the need to solve the complete inverse kinematics in order to arrive at a starting position. In general this transformation has many solutions, and it is difficult to add other constraints which may be important (e.g., reducing loading or avoiding obstacles). Using a robot of human-like shape constrains the possible solutions, and reduces the overall computational complexity of the task*(1, 3)*.

If the robot and human have a similar shape the robot will be able to better model what the instructor is doing. This knowledge will help the robot infer how the task looks from the instructor's perspective and what information it needs to perform the task itself.

Embodiment helps mapping between bodies. For the robot to be able to imitate the instructor, a mapping between the instructor's body and its own body must be established *(2)*. This task is greatly simplified if the robot and the instructor have a similar body shape. With a different morphology, not only is the mapping from human to robot more difficult, but the actual way the task is achieved is different. For example, consider the difference between a one and two armed robot performing a complex task. Having two arms completely changes how the action can be performed, since one arm can hold the object, while the other manipulates it, rather than having to clamp object.

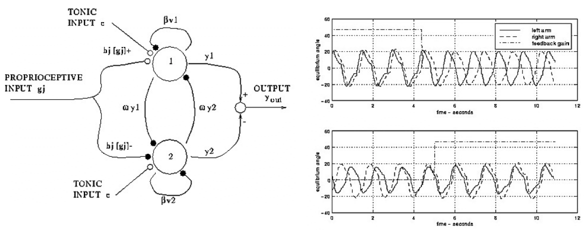

Figure 8: We have used neural oscillator circuits to exploit the natural dynamics of Cog's arm. The two consecutive traces on the right show the entrainment through the mechanical world of two arms being controlled by independent oscillators as they play with a slinky. When proprioception is switched off, entrainment decays but rapidly returns below when it proprioception is turned back on.

Embodiment allows intuitive interactions. Giving the robot a similar body shape also makes interaction with humans more intuitive *(6)*. The instructor can easily interpret the robots physical configuration, understand what is wrong or right, and assist accordingly. The instructor can also test options using his or her own body before instructing the robot.

Embodiment enables simple, robust low-level behaviors. Humans exploit the natural dynamics of their bodies and the tools that they manipulate. They respond the natural dynamics when manipulating objects (Turvey & Carello 1995), throwing objects (Bingham, Schmidt & Rosenblum 1989), and when walking (Alexander 1990, McGeer 1990). For the robot to effectively imitate these actions, it needs to have a similar shape and dynamics to a human body. For example, when throwing objects, humans exploit the spring-like properties of their arms, as well as the inter-segmental dynamical forces between the arm links. Attempting to use the same solution on a stiff robot with a different morphology would not produce efficient throwing. Our robots have special actuators which ensure that their dynamics are spring-like (Pratt & Williamson 1995).

Work in our lab has suggested that controlling robots by exploiting the natural dynamics of the arm-environment system can allow very simple controllers to perform complex tasks in a stable and robust manner (Williamson 1998). We have been using non-linear oscillators (figure 8) to excite the dynamics of robot arms and using feedback from the system dynamics to modify the oscillator outputs. Using these simple controllers (each equivalent to two biological

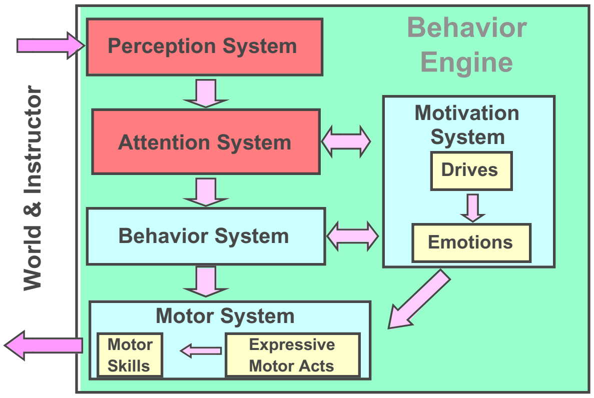

Figure 9: This is our high level system architecture for integrating perception, attention, behavior, motivation and motor systems. We propose fleshing out the details in a much more realistic way in order to enable a human to task the robot.

neurons), the arm has performed a variety of complex coordinated tasks such as crank turning and pumping a bicycle pump, as well as directly responding to system dynamics by tuning to the resonant frequency of driven pendulums. The arms have also been used for tasks such as throwing and hammering. The final solutions in all these cases are stable and robust to external perturbations because they exploit the natural dynamics of the system, rather than imposing a complete control structure. Having a physical robot and exploiting its natural dynamics gives simple and robust low-level behaviors which can generalize to a variety of tasks *(3, 5)*.

We have also found that using this approach results in solutions which are easy to learn. The main parameter to be learned is the starting posture of the arm, which in the above examples was taken from plausible human solutions, and which can be inferred from an instructor's demonstration.

In addition, since many tools and environments are designed for human rather than robot use (gas cap), a robot will be better able to use those tools if it has both a human morphology, and a similar dynamical structure.

3.IV Integration

The integration of multiple sensory modalities, physical degrees of freedom, and behavioral systems in a single robot allows the robot to imitate and interact with humans in a more sophisticated manner (figure 9). Integration is necessary at several levels, from the mapping of primitive sensors with primitive motor actions to the combination of goal directed behaviors and attention. The following subsections describe in detail how integration is an asset to our robotic systems.

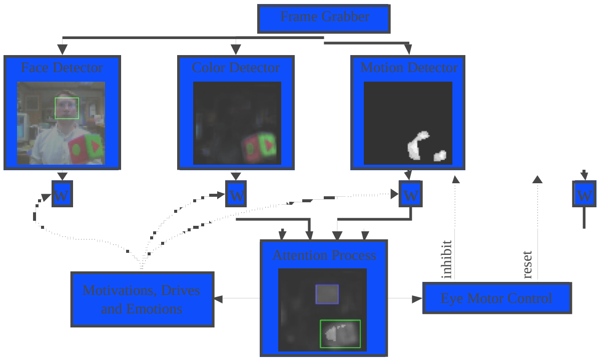

Figure 10: The attentional system of Kismet integrates multiple, bottom-up sensory feature detectors with top-down task-dependent motivational influences from Kismet's higher level control system.

Integration provides intuitive interactions. When two humans are communicating, the conversation is a complex mixture of signals in different sensory modalities: gesture, speech, eye contact, facial expression, and physical contact. Intuitive and natural communication between humans and robots is aided if the robot can both sense and understand these human signals, as well as produce these signals. A robot processing and integrating this multi-modal information will be capable of intuitive interaction *(6)*.

One example of this is auditory localization. When audible instructions are given to the robot, it should orient and look at the source of the sound. The fact that sound and visual motion are often correlated (lips and speech, clapping, etc.) is used by our robots to robustly learn the mapping between sound signals and localization directions.

Integration helps a robot know what to imitate. Finding the salient feature in a scene is easier as more sensory modalities are available. For example, in [**a**], when the instructor demonstrates the gas filling task, the robot must understand the saliency of the gas cap. As the instructor handles the cap different sensory modalities combine to indicate its saliency. There is motion near the cap, sound from the door opening, and there are gestures towards the cap. By exploiting the inputs from all of its modalities, the robot can ascertain the salient aspects of the demonstration *(1)*.

An example from our work is the current attentional system of Kismet shown in figure 10. Kismet only attends to objects that are significant relative to its current emotionally-driven goals. Kismet exploits various sensory modalities; it discerns color, motion, and can detect faces. These inputs flow through its attentional system to ultimately influence its motor system *(1)*.

Integration aids behavior generalization and action chaining. An integrated system allows behaviors and skills that have been learned in one set of modalities to be transfered to others more easily *(5)*. Also, a demonstrated sequence is usually communicated in multiple modalities which the robot must map to its own system requiring different modalities and sequence for its imitation *(2, 4)*.

For example, the mobile robot Yuppy (Spud's predecessor) generalizes its visually-driven behaviors to include triggers of different sensory cues. Yuppy is programmed to avoid stimuli with certain colors and shapes and then learns to associate the accompanying sounds alone with its avoidance behavior.

Integration makes engineering sense. Integration provides engineering robustness and efficiency. Robustness addresses failure recovery and and recognizing success because systems of different modalities serve as complements, backup and verification *(3)*. Efficiency results from exploiting signals collected from sensors or sent to motors that are well-suited to the job rather than those that are not naturally helpful. For example, in the earthquake scenario [**b**] both the visual and auditory inputs are necessary to overcome the background noise.

In our work on Cog, we have implemented both a vestibular-ocular reflex (VOR) and optokinetic response (OKR) for image stabilization. Our implementation integrates inertial and visual information to produce a stable image on which other processing can be performed. Removing the effect of base motion using visual processing alone would be exceedingly difficult and computationally expensive.

4 Conclusions

There are many challenging aspects inherent in building a humanoid robot which can interact naturally with humans. These range far beyond the existing well understood, but extremely challenging, problems of building an integrated humanoid which can locomote, navigate, and interact with objects. By decomposing the human interaction problems into issues and approaches, it is possible to identify appropriate modules that can be constructed towards the ultimate goal.

Acknowledgements

Many members of the humanoid robot group at the MIT Artificial Intelligence Laboratory contributed to this paper. They included Matthew Marjanovic, Juan Velasquez, and Charles Kemp. Funding for this research was provided by the Office of Naval Research.

References

Alexander, R. M. (1990), 'Three Uses for Springs in Legged Locomotion', *International Journal of Robotics Research* **9**(2), 53–61.

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

Bingham, G. P., Schmidt, R. C. & Rosenblum, L. D. (1989), 'Hefting for a maximum distance throw: A smart perceptual mechanism', *Journal of Experimental Psychology: Human Perception and Performance* **15**(3), 507–528.

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

Breazeal, C. & Scassellati, B. (2000), 'Infant-like Social Interactions between a Robot and a Human Caretaker', *Adaptive Behavior.* To appear.

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'.

Brooks, R. A. (1996), Prospects for Human Level Intelligence for Humanoid Robots, *in* 'Proceedings of the First International Symposium on HUmanoid RObots, HURO'96', Tokyo, Japan, pp. 17–24.

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.

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

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

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

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.

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

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

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

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

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

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

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

Turvey, M. T. & Carello, C. (1995), Dynamic Touch, *in* W. Epstein & S. Rogers, eds, 'Perception of Space and Motion', Academic Press, pp. 401–490.

Velasquez, J. (1998), When Robots Weep: A Mechanism for Emotional Memories, *in* 'Proceedings of th 1998 National Conference on Artificial Ingelligence, AAAI98', pp. 70–75.

Williamson, M. M. (1998), 'Neural Control of Rhythmic Arm Movements', *Neural Networks* **11**, 1379–1394.

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