Knowing What to Imitate and Knowing When You Succeed

Brian Scassellati
MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139, USA
scaz@ai.mit.edu

Abstract

If we are to build robots that can imitate the actions of a human instructor, the robotics community must address a variety of issues. In this paper, we examine two of these issues. First, how does the robot know which things it should imitate? Second, how does the robot know when its actions are an adequate imitation of the original? We further describe an on-going research effort to implement systems for a humanoid robot that address these issues.

1 Introduction

Humans (and other animals) acquire new skills from social interactions with others through direct tutelage, observational conditioning, goal emulation, imitation, and other methods (Galef 1988). These social learning skills provide a powerful mechanism for children to acquire skills and knowledge from their parents, other adults, and other children. In particular, imitation is an extremely powerful mechanism for social learning which has received a great deal of interest from researchers in the fields of animal behavior and child development.

Similarly, social interaction can be a powerful way for transferring important skills, tasks, and information to a robot. A socially competent robot could take advantage of the same sorts of social learning and teaching scenarios that humans readily use. From an engineering perspective, a robot that could imitate the actions of a human would provide a simple and effective means for the human to specify a task to the robot and for the robot to acquire new skills without any additional programming. From a computer science perspective, imitation provides a means for biasing learning and interaction. From a developmental psychology perspective, building systems that learn through imitation allows us to investigate a minimal set of competencies necessary for social learning. We can further speculate that constructing an artificial system may provide useful information about the nature of imitative skills in humans (or other animals).

Research into social robotics began with studies of the collective behavior of groups of similar mobile robots including flocking, following, and homing through very simple communication channels (Balch & Arkin 1994, Matarić 1994). Initial studies of imitation focused on allowing one robot to imitate the navigation acts of a second robot using simple perception (proximity sensors and IR following) through mazes (Hayes & Demiris 1994) or an unknown landscape (Dautenhahn 1995). Other work in social learning for autonomous robots addressed learning inter-personal communication protocols between similar robots (Steels 1996), and between robots with similar morphology but which differ in scale (Billard & Dautenhahn 1997). Matarić, Williamson, Demiris & Mohan (1998) has addressed imitation by studying human imitation of complex motor tasks (Matarić & Pomplun 1998), as well as implementing some imitation tasks on a simulated humanoid (Demiris & Matarić 1998).

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

Even if we were to construct a system which had all of the requisite precursor skills, the act of imitation also presents its own unique set of research questions. Consider the following example: The robot is observing a man opening a glass jar. The man approaches the robot and places the jar on a table near the robot. The man rubs his hands together and then sets himself to removing the lid from the jar. He grasps the glass jar in one hand and the lid in the other and begins to unscrew the lid. While he is opening the jar, he pauses to wipe his brow, and glances at the robot to see what it is doing. He then resumes opening the jar.

When observing this scene, how does the robot determine that this is an appropriate action to imitate? How does the robot separate the sensory scene into the components that are relevant to the task (such as the jar) and which are not relevant (such as the table or the man's clothing). Further, how does the robot determine which actions (such as grasping the lid) are relevant to the task, which actions (such as the glance at the robot) have social significance but are not part of the task, and which actions (such as wiping one's brow) are irrelevant to the task? Once the robot has recognized the appropriate actions, how does that perceptual information (such as the movement in the visual scene resulting from twisting the lid) map into motor actions that the robot is capable of performing (an arm movement that would rotate the wrist and elbow)? How does the robot combine each of the individual actions that it observes into a coherent and flexible sequence of behaviors, allowing for omissions and additions under appropriate circumstances? How does the robot know when it should attempt to perform the action that it has observed? On what types of sensory stimuli should the observed action be attempted? After observing the man with the jar, should the robot attempt to open jars of differing colors and shapes? After the robot attempts the action, how does it know that it has been successful? How does the robot know the intended result of the action? How does the robot evaluate its own attempt, and if inaccurate, how does the robot better its subsequent attempts? After a successful action has been performed, how is this action generalized for new target objects, situations, and conditions?

Each of these questions is a complex research problem which the robotics community has only begun to address. In this paper, we will examine two of these issues: "How does the robot know what to imitate?" and "How does the robot evaluate its performance?" To simplify our discussion of these issues (not to mention the implementation), we start with the slightly easier problem of learning to imitate from a helpful instructor (learning through direct tutelage) rather than the more difficult problem of learning to imitate under adversarial or indifferent conditions. While this assumption does limit the generality of our discussions, it does more accurately represent the learning environment of human infants (who constantly benefit from the help and encouragement of their caregivers).

In the following sections, we will focus on the research issues that arise in attempting to build systems that address these two research issues for a humanoid robot. Sections 2 and 3 discuss some of the difficulties that arise in addressing these issues and begin to describe the research methodology that we have applied to these problems. Section 4 describes two robotic platforms (an upper-torso humanoid robot and an active vision system with expressive displays) that we are using in building systems that can imitate. Section 5 describes our current progress on building segmentation and attentional systems within a social framework, and on building systems that recognize social cues.

2 How do you know 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. When attempting to imitate another individual, how does the robot determine which aspects of its sensory environment are relevant? Assuming the robot has identified the human instructor, how does it determine which of the instructor's actions are relevant to the task and which are circumstantial? For example, to imitate placing a lid on a jar, the robot must segment the scene into salient objects (such as the instructor's hand, the lid, and the jar) and actions (the instructor's moving hand twisting the cap and the instructor's head turning toward the robot). The robot must determine which of these objects and events are necessary to the task at hand (such as the jar and the movement of the instructor's elbow), which events and actions are important to the instructional process but not to the task itself (such as the movement of the instructor's head), and which are inconsequential (such as the instructor wiping his brow). The robot must also determine to what extent each action must be imitated. For example, in removing the lid from a jar, the movement of the instructor's hand is a critical part of the task while the instructor's posture is not.

In addressing this issue, four aspects of our research methodology have been critical: capitalizing on innate social interactions with the instructor, constructing a developmental progression of skills which build gracefully toward imitation, exploiting the advantages of the robot's physical embodiment, and leveraging the integration of multiple sensory and motor modalities to provide robust and flexible behaviors.[1]

2.1 Using social cues to determine saliency

Fundamental social cues (such as gaze direction) can be used by a robot to determine the important features of a task. Human instructors naturally attend to the key aspects of a task when demonstrating that task. For example, when opening the jar, the instructor will naturally look at the lid as he grasps it and at his own hand while twisting off the lid. By directing its own attention to the object of the instructor'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 instructor; the instructor glances over his shoulder to confirm that the robot is looking in the right

[1] These research methods are more fully explored for other aspects of humanoid robotics in Brooks, Breazeal (Ferrell), Irie, Kemp, Marjanović, Scassellati & Williamson (1998).

place. If this is not the case, then the instructor can actively direct the robot's attention to the jar, perhaps by pointing to it or tapping on it. In general, knowledge of basic social cues is necessary to distinguish acts of communication from acts directly related to the task being taught.

2.2 Developmental progressions limit complexity

Humans are not born with complete reasoning systems, complete motor systems, or even complete sensory systems. Instead, they undergo a process of development in which 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 correctly without any experience in the world. Human development provides us with insight into how complex behaviors and skills (such as manipulating an object or perceiving where the instructor's attention is focused) can be broken down into simpler behaviors. Acquired 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 enable our robots to learn increasingly sophisticated behaviors.

Systems that follow human-like developmental paths allow increasingly more complex skills and competencies to be layered on top of simpler competencies. A developmental approach keeps the complexity of 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 jar from one place to another without getting lost in the details of the action.

2.3 Physical morphology constrains perception

If the robot and human have a similar shape, the space of possible actions that the robot must select from constrains the perceptual task. For example, if the robot observes a man doing something to a jar with his hands, the robot can discard any potential perceptions which do not match to actions that it is capable of performing with its own hands. Additionally, the position of the instructor's arm serves as a guideline for an initial configuration for the robot's arm. 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.

2.4 Cross-modal perceptual constraints

Finding the salient features in a social interaction becomes easier as more sensory modalities are available. For example, when the instructor unscrews the lid from the jar, sensory cues from the visual system (motion) and the auditory system (the sound of the lid being unscrewed) occur at the same time and in the same spatial location. These correlations can be exploited to better refine the perceptions of each individual modality. For example, the visual motion cues can aid in the localization of the auditory stimulus, which may in turn lead to a better audio segmentation.

2.5 Applying these methodologies

Our research on this issue has focused on two areas: recognizing inherent saliency in objects and recognizing objects and actions that are salient as a result of the attentional state of the instructor. To recognize inherent object saliency, we have been constructing attentional and perceptual systems that combine information on visual motion, innate perceptual classifiers such as face detectors, color saliency, depth segmentation, and auditory information with a habituation mechanism and a motivational and emotional model. This attentional system will allow the robot to selectively direct computational resources and exploratory behaviors toward objects in the environment that have inherent saliency. While these attentional systems provide context-dependent saliency information, we also utilize the observed attentional states of the human instructor as a means of determining which actions and objects are relevant. We already have perceptual systems that allow us to detect faces, move the eyes to the detected face, and obtain a high-resolution image of the instructor's eyes (Scassellati 1998). We are currently working on utilizing information on the location of the pupil, the

angle of gaze, the orientation of the head, and body posture to determine the object of the instructor's attention. This emphasis on joint reference is part of a larger project to build a "theory of mind" for the robot, which would allow it to attribute beliefs, desires, and intentions to the instructor (Scassellati 1999).

3 How do you know when you have been successful?

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 unscrew the lid of a jar, has the robot been successful if it rotates the lid but leaves the lid on the jar? Is the robot successful if it removes the lid but empties the contents of the jar onto the floor? In each 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. The instructor continually modifies the way he performs the task, perhaps exaggerating those aspects that the student performed inadequately, in an effort to refine the student's subsequent performance. By repeatedly responding to the 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.

Social interaction plays 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 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 behaviors. 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.

The physical morphology of the robot can also assist in evaluating success. If the robot's morphology is similar to the instructor's, then the robot is likely to have similar failure modes. This potentially allows the robot to characterize some of its own failures by observing the failures of the instructor. If the robot watches the instructor having difficulty opening the jar when his elbows are close together, the robot may be able to extrapolate that it too will fail without sufficient leverage. A similar morphology also allows the instructor to more easily identify and correct errors from the robot. If the robot's arms are too close together when attempting to open the jar, the instructor's knowledge about his own body will assist him in providing feedback to the robot.

With our robots, we plan on using joint reference as a cue for iterative refinement. We have also planned an implementation of an auditory system capable of detecting prosody (including pitch, tempo, and tone of voice) as a secondary signal for obtaining feedback on which actions have been successfully executed and which have not. We are also currently engaged in building a facial expression recognition system which would allow the robot to obtain feedback directly from the facial expressions of the instructor.

4 Robotic platforms

Our work with imitation has focused on two platforms: an upper-torso humanoid robot called Cog and an active vision system enhanced with facial features called Kismet (see Figure 1). Both of these robots have been constructed in part to investigate how to build intelligent robotic systems by following a developmental progression of skills similar to that observed in human development (Brooks & Stein 1994, Brooks et al. 1998). In the past two years, a basic repertoire of perceptual capabilities and sensorymotor skills have been implemented on these robots (see Brooks, Breazeal, Marjanovic, Scassellati & Williamson (1999) for a review).

Cog approximates a human being from the waist up with twenty-one degrees-of-freedom (DOF) and a variety of sensory systems. The physical structure of the robot, with movable torso, arms, neck and eyes gives it humanlike motion, while the sensory systems (visual, auditory, vestibular, and proprioceptive) provide rich information about the robot and its immediate environment. These together present many opportunities for interaction between the robot and humans.

The robot Kismet is based on the same active vision system used on Cog. In addition to the three degrees-of-freedom in the eyes, Kismet also has one degree-of-freedom in the neck and eleven degrees-of-freedom in facial expressions, including eyebrows (each with two degrees-of-freedom: lift and arch), ears (each with two degrees-of-freedom: lift and rotate), eyelids (each with one degree of freedom: open/close), and a mouth (with one de-

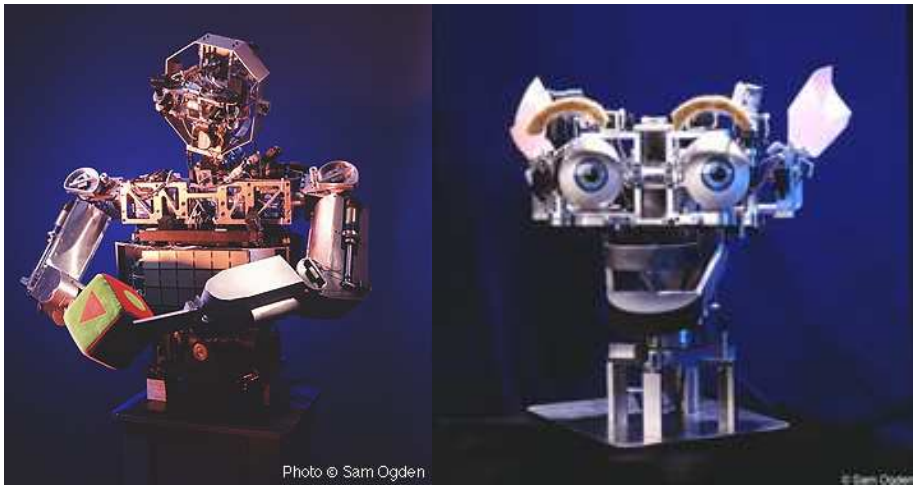

Figure 1: At left, Cog, an upper-torso humanoid robot with twenty-one degrees of freedom and auditory, visual, vestibular, and kinesthetic sensory systems. At right, Kismet, an active vision system with facial expressions.

gree of freedom: open/close). The robot is able to show expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise in response to perceptual stimuli (Breazeal & Scassellati 1998).

By focusing on robotic platforms that are anthropomorphic, we simplify the problems of social interaction in three ways. First, it allows for a simple and natural means of interaction. People already know how to provide the robot with appropriate feedback, how to attract its attention, and can guess what capabilities it might possess. Second, the responses of the robot can be easily identified and interpreted by a naive observer. Third, by having a similar body structure, the problem of mapping observed actions onto the robot's own body is simplified.

5 Imitation behaviors

Many of the precursor skills that we have described above have already been implemented on our robots, including perceptual abilities like face detection, motion detection, and disparity filtering, visual-motor skills such as orienting behaviors, smooth tracking, and gaze stabilization reflexes, as well as basic attentional and motivational systems. Many more of these skills are currently under development (especially an advanced attentional system, gaze direction identification, and facial gesture recognition). In this section, we focus on two sets of skills which directly impact the problems of determining what to imitate and knowing when you have succeeded: an attentional system that integrates both inherent object properties with high-level goal-oriented knowledge to determine saliency and a system which can identify the attentional states of the instructor in order to provide better saliency and evaluate performance.

5.1 Attentional system

We have constructed an attentional system for the robot Kismet based upon Wolfe's model of human visual attention and visual search (Wolfe 1994). This model integrates evidence from Treisman (1985), Julesz & Krose (1988), and others to construct a flexible model of human visual search behavior. In Wolfe's model, low-level perceptual inputs are combined with high-level influences from motivations and task demands.

The attention system attributes saliency to stimuli that exhibit certain low-level, pre-attentive feature properties which human infants find interesting. For example, a four-month-old infant is more likely to look at a moving object than a static one, or a face-like object than one that has similar, but jumbled, features (Fagan 1976). To mimic the preferences of human infants, Kismet's attention system combines three basic feature detectors: face finding, motion detection, and color saliency analysis. The face finding system recognizes frontal views of faces within approximately six feet of the robot under a variety of lighting conditions (Scassellati 1998). The motion detection module uses temporal differencing and region growing to obtain bounding boxes of moving objects (Breazeal & Scassellati 1998). Color content is computed using an opponent-process model that identifies saturated areas of red, green, blue, and yellow (Breazeal & Scassellati 1999). All of these systems operate at speeds that are amenable to social interaction (20-30Hz).

The attention process constructs a linear combination of the input feature detectors and a time-decayed Gaussian field which represents habituation effects (see Figure 2). Top-down influences from motivational, emotional, and task constraints can influence the attention selection process by changing the relative contributions of the input feature detectors. For example, if the robot has become bored and lonely, the weight of the face detector can be increased to preferentially bias the robot to attend

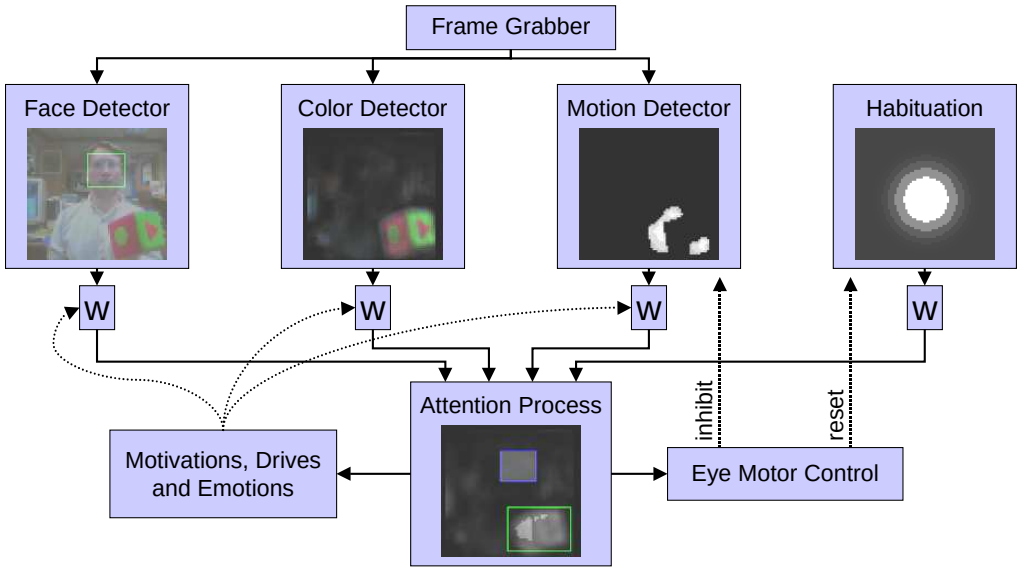

Figure 2: The attentional system of Kismet integrates multiple, bottom-up sensory feature detectors with top-down task-dependent motivational influences from Kismet's higher level control system.

to faces. This combination of low-level inherent saliency of perception with state-based selection allows for attentional choices that are goal-driven while remaining opportunistic with respect to the incoming perceptual stimuli (Breazeal & Scassellati 1999).

We have also extended this attentional system to regulating the rate and intensity of face-to-face interactions with a human caretaker (Breazeal & Scassellati 1998). Perceptual stimuli that are selected by the attention process are classified into *social* stimuli (i.e. people, which move and have faces) which satisfy a drive to be social and *non-social* stimuli (i.e. toys, which move and are colorful) which satisfy a drive to be stimulated by other things in the environment. Just as infants manipulate their parents (Trevarthen 1979), 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 and turns to look away. These social cues are unconsciously interpreted by the instructor, who modifies his behavior to maintain the interaction (see also the submission by Breazeal in this volume).

This attentional system is the basis of a system that can determine which objects and events are relevant based upon the current task constraints, internal environment, and external stimuli. One addition that would greatly enhance the robot's ability to detect important objects and events is the ability to recognize the attentional state of the instructor.

5.2 Detecting attentional states

One critical milestone in a human 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, Bruner & Ross 1976).

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 sup-

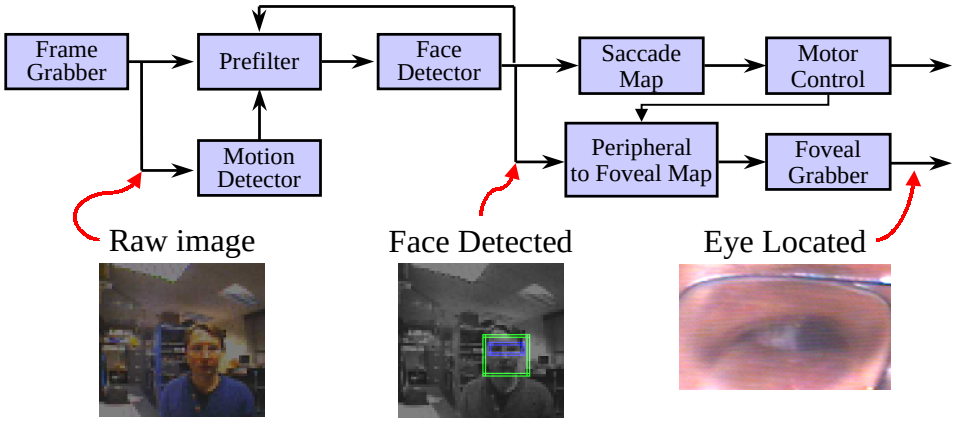

Figure 3: This is our existing system for face and eye detection in cluttered scenes. We are currently extending this system to recognize and interpolate gaze direction.

port from developmental psychology, neuroscience, and abnormal psychology, and is consistent with evidence from evolutionary studies of the development of joint attention behaviors (such as Povinelli & Preuss (1995) and Hauser (1996)). This decomposition is described in detail by Scassellati (1999). In this section, we give a brief overview of our progress on building systems that respond to attentional states.

In normal development, infants are first capable of recognizing and maintaining eye contact. This ability develops over the course of the first 18 months of life to allow the infant to extrapolate the angle of gaze and engage in gaze-following behaviors (a rapid alternation between looking at the eyes of the individual and looking at the distal object of their attention). This simplest form of joint attention is believed to be critical for social scaffolding(Thelen & Smith 1994), development of theory of mind(Baron-Cohen 1995), and providing shared meaning for learning language (Wood et al. 1976). As gaze following matures, the infant begins to exhibit a second form of joint reference, declarative pointing. Declarative pointing is characterized by an extended arm and index finger designed to draw attention to a distal object. Declarative pointing differs from an earlier from imperative pointing (a gesture used to obtain an object that is out of reach by pointing at that object) in that it is not necessarily a request for an object; children often use declarative pointing to draw attention to objects that are clearly outside their reach, such as the sun or an airplane passing overhead.

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 has been identified, the robot saccades to that target to obtain a high resolution image of the eye (see Figure 3). This technique has been successful at locating and extracting sub-images that contain eyes under a variety of conditions and from many different individuals. We are currently extending this system to identify pupil location and head orientation in order to detect the angle of gaze.

Implementing imperative pointing is accomplished by implementing the more generic task of reaching to a visual target. Following the developmental path that Diamond (1990) demonstrated in infants between five and twelve months of age, Marjanović, Scassellati & Williamson (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. The task of recognizing a declarative pointing gesture can be seen as the application of the geometric and representational mechanisms for gaze following to a new initial stimulus. Instead of extrapolating from the vector formed by the angle of gaze to achieve a distal object, we extrapolate the vector formed by the position of the arm with respect to the body. This requires a rudimentary gesture recognition system, but otherwise utilizes the same mechanisms.

Recognizing the attentional states of the instructor assists in knowing what to imitate and in evaluating success. By monitoring the instructor, the robot can obtain powerful cues about the objects that are important to a task. The robot can also evaluate its own performance based on the emotive, postural, and attentional states of the instructor. 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.

5.3 Simple head nod imitation

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 4). 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.

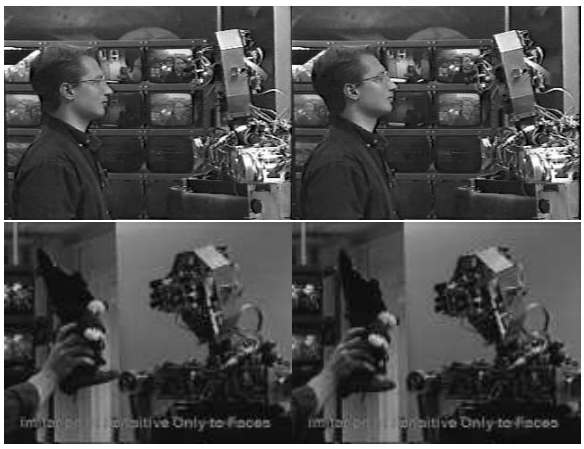

Figure 4: Images captured from a videotape of the robot imitating head nods. The two images at left show the robot imitating head nods from a human caretaker. The output of the face detector is used to drive fixed yes/no nodding responses in the robot. The face detector also picks out the face from stuffed animals, and will also mimic their actions (right images). The original video clips are available at `http://www.ai.mit.edu/projects/cog/`.

6 Conclusion

Building a system that can imitate the actions of a human instructor is an extremely complex task with many open research questions. In this paper, we have examined two questions relating to imitation: "How do you know what to imitate" and "How do you know when you have it right?" We have begun to build systems for a pair of anthropomorphic robots that address some of the issues that these two questions raise. An attentional system that is sensitive both to inherent object properties and high-level task constraints assists in recognizing the salient parts of an action. These results can be augmented by examining the attentional states of the instructor, a technique which can also be used to obtain evaluations from the instructor and selectively improve components of an imitative act.

Acknowledgments

Support for this project is provided in part by an ONR/ARPA Vision MURI Grant (No. N00014-95-1-0600). The author wishes to thank Rod Brooks, Cynthia Breazeal, and Una-May O'Reilly for their comments and suggestions on pieces of this work.

References

Balch, R. & Arkin, R. (1994), 'Communication in Reactive Multiagent Robotic Systems', *Autonomous Robots*.

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

Billard, A. & Dautenhahn, K. (1997), Grounding Communication in Situated, Social Robots, Technical report, University of Manchester.

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

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

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

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

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

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

Demiris, J. & Matarić, M. J. (1998), Perceptuo-Motor Primitives in Imitation, *in* 'Working Notes, Autonomous Agents '98 Workshop on Agents in Interaction - Acquiring Competence', Minneapolis/St Paul.

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.

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

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

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

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

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

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

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

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.

Matarić, M. (1994), Reward functions for accelerated learning, *in* 'Proceedings of the eleventh international conference on machine learning', New Brunswick, NJ, pp. 181–189.

Matarić, M. J. & Pomplun, M. (1998), 'Fixation Behavior in Observation and Imitation of Human Movement', *Cognitive Brain Research* **7**(2), 191–202.

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

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

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.

Steels, L. (1996), Emergent Adaptive Lexicons, *in* 'Proceedings of the fourth international conference on simulation of adaptive behavior'.

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

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

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.

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

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