Infant-like Social Interactions between a Robot and a Human Caretaker

Cynthia Breazeal (Ferrell)
Brian Scassellati
Massachusetts Institute of Technology
Artificial Intelligence Laboratory
545 Technology Square, Room 938
Cambridge, MA 02139 USA
(617) 253-7593
fax: (617) 253-0039
email: ferrell@ai.mit.edu, scaz@ai.mit.edu

Abstract

This paper presents an autonomous robot designed to interact socially with human "parents". A human infant's emotions and drives play an important role in generating meaningful interactions with the caretaker, regulating these interactions to maintain an environment suitable for the learning process, and assisting the caretaker in satisfying the infant's drives. For our purposes, the ability to regulate how intensely the caretaker engages the robot is vital to successful learning in a social context.

To achieve a similar interaction dynamic, we present a general framework that integrates perception, attention, drives, emotions, behavior selection, and motor acts. We then present a specific implementation of this architecture which enables the robot to perceive both salient social stimuli (faces) and salient non-social stimuli (motion). The robot responds with expressive displays which reflect an ever-changing motivational state and which give the human cues on how to satisfy the robot's drives while neither over-whelming nor under-stimulating the robot. Results from a series of experiments are presented where a human engages the robot in either direct face-to-face exchanges or with a toy. We believe this work is an important step toward realizing autonomous robots that can engage in meaningful bi-directional social interactions with humans.

keywords: Human-robot interaction, social agents, emotional agents

shortened title: A Social Infant-like Robot

1 Introduction

Social robotics has generally concentrated on the behavior of groups of robots performing behaviors such as flocking, foraging or dispersion (Mataric 1995, Balch & Arkin 1994) or on paired robot-robot interactions such as imitation (Billard & Dautenhahn 1997). Our work focuses not on robot-robot interactions, but rather on the construction of robots that engage in meaningful social exchanges with humans. By doing so, it is possible to have a socially sophisticated human assist the robot in acquiring more sophisticated communication skills and helping it learn the meaning these acts have for others. Our approach is inspired by the way infants learn to communicate with adults. Specifically, the mode of social interaction is that of a caretaker-infant dyad where a human acts as the caretaker for the robot.

An infant's emotions and drives play an important role in generating meaningful interactions with the caretaker (Bullowa 1979). These interactions constitute learning episodes for new communication behaviors. In particular, the infant is strongly biased to learn communication skills that result in having the caretaker satisfy the infant's drives (Halliday 1975). The infant's emotional responses provide important cues which the caretaker uses to assess how to satiate the infant's drives, and how to carefully regulate the complexity of the interaction. The former is critical for the infant to learn how its actions influence the caretaker, and the later is critical for establishing and maintaining a suitable learning environment for the infant.

This paper presents the first stages of this long term endeavor. We describe a framework which integrates perception, attention, drives, emotions, behavior arbitration, and expressive acts. We have used this framework to implement an autonomous robot that is specialized for regulating the intensity of social interaction – an elaborated version from that presented in (Breazeal(Ferrell) 1998). We concentrate on the design specification of the perceptual and motivational systems because of the critical role they serve in this dynamic process for infants. Other work in progress focuses on the construction of shared attentional systems that allow the infant and the caretaker to ground learning in perceptual episodes (Scassellati 1996, Scassellati 1998*c*).

Although we do not claim that this system parallels infants exactly, its design is heavily inspired by the role motivations and facial expressions play in maintaining an appropriate level of stimulation during social interaction with adults. This is a critical skill for the kinds of social learning that mothers and infants engage in, for it helps the mother tune her actions so that they are appropriate for the infant. For our purposes, the context for learning involves social exchanges where the robot learns how to manipulate the caretaker into satisfying its internal drives. Ultimately, the communication skills targeted for learning are those exhibited by infants such as turn taking, shared attention, and pre-linguistic vocalizations exhibiting shared meaning with the caretaker.

This paper is organized as follows: first we discuss the numerous roles motivations play in natural systems—particularly as it applies to behavior selection, regulating the intensity of social interactions, and learning in a social context. Next we describe a robot called Kismet that we have designed and built to provide emotional feedback to the caretaker through facial expressions. We then present a framework for the design of the behavior engine which integrates perception, motivation (drives and emotions), attention, behavior, and motor skills (expressive or task based). Particular detail is given for the design of the perceptual and motivational systems. After we illustrate these ideas with a specific implementation on a physical robot, we present the results of some early experiments where a human engages the robot in face-to-face social exchanges. Finally, we discuss planned extensions to the existing system.

2 The Role of Motivations in Social Interaction

Motivations encompass drives, emotions, and pain. Motivations play several important roles for both arbitrating and learning behavior. For our purposes, we are interested in how they influence behavior selection, regulate social interactions, and promote learning in a social context.

2.1 Behavior Selection:

In ethology, much of the work in motivation theory tries to explain how animals engage in appropriate behaviors at the appropriate time to promote survival (Tinbergen 1951, Lorenz 1973). For animals, internal drives influence which behavior the animal pursues, for example, feeding, foraging, or sleeping. Furthermore, depending on the intensity of the drives, the same sensory stimulus may result in very different behavior. For example, a dog will respond differently to a bone when it is hungry than when it is fleeing from danger.

It is also well accepted that animals learn things that facilitate the achievement of biologically significant goals. Work in ethology has argued that motivations provide an impetus for this. In particular, the motivational system provides a reinforcement signal that guides what the animal learns and in what context. When an animal has a strong drive that it is trying to satisfy, it is primed to learn behaviors that directly act to satiate that drive. For this reason, it is much easier to train a hungry animal with a food reward than a satiated one (Lorenz 1973).

For a robot, an important function of the motivation system is to regulate behavior selection so that the observable behavior appears coherent, appropriately persistent, and relevant given the internal state of the robot and the external state of the environment. The responsibility for this function falls largely under the drive system of the robot. Other work in autonomous agent research has used drives in a similar manner (Maes 1990, Arkin 1988, McFarland & Bosser 1993, Steels 1995). Drives are also necessary for establishing the context for learning as well as providing a reinforcing signal. Blumberg (1996) used motivations (called *internal variables*) in this way to implement operant conditioning so that human user could train an animated dog new tricks.

2.2 Regulating Interaction:

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

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

2.3 Learning in a Social Context:

The use of emotional expressions and gestures facilitates and biases learning during social exchanges. Parents take an active role in shaping and guiding how and what infants learn by means of *scaffolding*. As the word implies, the parent provides a supportive framework for the infant by manipulating the infant's interactions with the environment to foster novel abilities. Commonly, scaffolding involves reducing distractions, marking the task's critical attributes, reducing the number of degrees of freedom in the target task, providing ongoing reinforcement through expressive displays of face and voice, and enabling the subject to experience the end or outcome of a sequence of activity before the infant is cognitively or physically able of seeking and attaining it for himself (Wood, Bruner & Ross 1976). The emotive cues the parent receives during social exchanges serve as feedback so the parent can adjust the nature and intensity of the structured learning episode to maintain a suitable learning environment where the infant is neither bored nor over-whelmed.

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

For example, early cries of an infant elicit various care-giving responses from his mother depending upon how she initially interprets these cries and how the infant responds to her mothering acts. Over time, the infant and mother converge on specific meanings for different kinds of cries. Gradually the infant uses subtly different cries (i.e., cries of distress, cries for attention, cries of pain, cries of fear) to elicit different responses from his mother. The mother reinforces the shared meaning of the cries by responding in consistent ways to the subtle variations. Evidence of this phenomena exists where mother-infant pairs develop communication protocols different from those of other mother-infant pairs (Bullowa 1979).

Combining these ideas one can design a robot that is biased to learn how its emotive acts influence the caretaker in order to satisfy its own drives. Toward this end, we endow the robot with a motivational system that works to maintain its drives within homeostatic bounds and motivates the robot to learn behaviors that satiate them. Further, we provide the robot with a set of emotive expressions that are easily interpreted by a naive observer as analogues of the types of emotive expressions that human infants display. This allows the caretaker to observe the robot's emotive expressions and interpret them as communicative acts. She assumes the robot is trying to tell her which of its needs must be tended to, and she acts accordingly. This establishes the requisite routine interactions for the robot to learn how its emotive acts influence the behavior of the caretaker, which ultimately serves to satiate the robot's own drives.

This section has argued that motivations should play a significant role in determining the robot's behavior, how it interacts with the caretaker, and what it can learn during social exchanges. With these long term challenges in mind, an important pre-requisite function for the robot's motivational system is not only to establish appropriate interactions with the human, but to also to regulate the interaction intensity so that the robot can learn without being over-whelmed or under-stimulated. When designed properly, the interaction among the robot's drives, emotions, and expressions provide appropriate cues for the caretaker so that she knows whether to change the activity itself or to modify its intensity. By doing so, both parties can modify both their own behavior and the behavior of the other in order to maintain an interaction that the robot can learn from and use to satisfy its drives.

3 Kismet's Hardware

To explore these ideas, we have constructed a robot with capabilities for emotive facial expressions, shown in figure 1. It consists of an active stereo vision system (described in (Scassellati 1998*a*)) embellished with facial features for emotive expression. Currently, these facial features include eyebrows (each with two degrees-of-freedom: lift and arch), ears (each with two degrees-of-freedom: lift and rotate), eyelids (each with one degree of freedom: open/close), and a mouth (with one degree of freedom: open/close). The robot is able to show expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise (shown in figure 2) which are easily interpreted by an untrained human observer.

Similar to other active vision systems (Sharkey, Murray, Vandevelde, Reid & McLauchlan 1993, Coombs 1992), there are three degrees of freedom; each eye has an independent vertical axis of rotation (pan) and the eyes share a joint horizontal axis of rotation (tilt). Each eyeball has a color CCD camera embedded within it having a 5.6 mm focal length. Although this limits the field of view, most social interactions require a high acuity central area to capture the details of face-to-face interaction. However, infants have poor visual acuity which restricts their visual attention to about two feet away – typically the distance to their mother's face when the infant is being held (Goldstein 1989). [1] This choice of camera is a balance between the need for high resolution and the need for a wide low-acuity field of view.

The active vision platform is attached to a parallel network of digital signal processors (Texas Instruments TMS320C40), as shown in figure 3. The DSP network serves as the sensory processing engine and implements the bulk of the robot's perception and attention systems. Each node in the network contains one processor with the option for more specialized hardware for capturing images, performing convolution quickly, or displaying images to a VGA display. Nodes may be connected with arbitrary bi-directional hardware connections, and distant nodes may communicate through virtual connections. Each camera is attached to its own frame grabber, which can transmit captured images to connected nodes.

[1] For example, at one month the infant has a visual acuity between 20/400 to 20/600.

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

4 A Framework for Designing Behavior Engines

A framework for how the motivational system interacts with and is expressed through behavior is shown in figure 4. The organization and operation of this framework is heavily influenced by concepts from psychology, ethology, and developmental psychology, as well as the applications of these fields to robotics as outlined in Brooks, Ferrell, Irie, Kemp, Marjanovic, Scassellati & Williamson (1998). The system architecture consists of five subsystems: the *perception system*, the *motivation system*, the *attention system*, the *behavior system*, and the *motor system*, an elaborated version from that presented in previous work (Breazeal(Ferrell) 1998). The perception system extracts salient features from the world, the motivation system maintains internal state in the form of `drives` and `emotions`, the attention system determines saliency based upon perception and motivation, the behavior system implements various types of behaviors as conceptualized by Tinbergen (1951) and Lorenz (1973), and the motor system realizes these behaviors as facial expressions and other motor skills.

The overall system is implemented as an agent-based architecture similar to that of (Blumberg 1996, Maes 1990, Brooks 1986, Minsky 1988). For this implementation, the basic computational process is modeled as a transducer. Each drive, emotion, behavior, percept, and facial expression is modeled as a separate transducer process specifically tailored for its role in the overall system architecture. The activation energy x of a transducer is computed by the equation: $x = (\sum_{n}^{j=1} w_j \cdot i_j) + b$ where i_j are inputs, w_j are weights, b is the bias, and n is the number of inputs. The weights can be either positive or negative; a positive weight corresponds to an excitatory connection and a negative weight corresponds to an inhibitory connection. The process is *active* when its activation level exceeds an *activation threshold*. When active, the process may perform some special computation, send output messages to connected processes, spread some of its activation energy to connected units, and/or express itself through behavior.

4.1 The Perception System

The responsibility of the perception system is to convert raw sensory stimuli into meaningful information to guide behavior. For this system, visual images are processed for both salient social stimuli (faces) and salient non-social stimuli (motion). These processed images feed a "face" percept and a "non-face" percept, each of which is modeled by a transducer. The intensity values are used to guide the robot's behavior – the robot responds in a manner to keep the "face" and "non-face" percepts within a desired intensity range. The design of the perceptual system is described in detail in section 5.

4.2 The Motivation System

The motivation system consists of two related subsystems, one which implements `drives` and a second which implements `emotions` and `expressive states`.[2] The `drives` serve as an internal representation of the robot's agenda, while the `emotions` and `expressive states` reflect how well the robot is achieving that agenda.

Motivations establish the nature of a creature by defining its needs and influencing how and when it acts to satisfy them (Lorenz 1973, Tinbergen 1951). The "nature" of this robot is to learn in a social environment. All `drives`, `emotions`, and behaviors are organized such that the robot is in a state of homeostatic balance when it is functioning adeptly and is in an environment that affords high learning potential. This entails that the robot be motivated to engage in appropriate interactions with its environment (including the caretaker) and that it is neither under-whelmed or over-whelmed by these interactions.

4.2.1 Drives:

The robot's `drives` serve three purposes. First, they influence behavior selection by preferentially passing activation to some behaviors over others. Second, they influence the emotive state of the robot by passing activation energy to the `emotion` processes. Since the robot's expressions reflect its emotive state, the `drives` indirectly control the expressive cues the robot displays to the caretaker. Third, they provide a learning context; the robot learns skills that serve to satisfy its `drives`.

The design of the robot's `drives` subsystem is heavily inspired by ethological views (Lorenz 1973, Tinbergen 1951). One distinguishing feature of drives is their temporally cyclic behavior. That is, given no stimulation, a drive will tend to increase in intensity unless it is satiated. For instance, an animal's hunger level or need to sleep follows a cyclical pattern.

Another distinguishing feature of drives are their homeostatic nature. For animals to survive, they must maintain a variety of critical parameters (such as temperature, energy level, amount of fluids, etc.) within a bounded range. As such, the `drives` of the robot change in intensity to reflect the ongoing needs of the robot and the urgency for tending to them. There is a desired operational point for each `drive` and an acceptable bounds of operation around that point. We call this range the *homeostatic regime*. As long as a `drive` is within the homeostatic regime, the robot's "needs" are being adequately met.

For this robot, each `drive` is modeled as a separate process with a temporal input to implement its cyclic behavior. The activation energy of each `drive` ranges between $[-max, +max]$, where the magnitude of the `drive` represents its intensity. For a given `drive` level, a large positive magnitude corresponds to being under-stimulated by the environment, whereas a large negative magnitude corresponds to being overstimulated by the environment. In general, each `drive` is partitioned into three regimes: an *under-whelmed regime*, an *over-whelmed regime*, and a *homeostatic regime*.

[2] As a convention, we will use the boldface to distinguish parts of the architecture of this particular system from the general uses of those words. In this case, "`drives`" refers to the particular computational processes that are active in the system, while "drives" refers to the general uses of that word.

4.2.2 Emotions and Expressive States:

The `emotions` of the robot serve two functions. First, they influence the emotive expression of the robot by passing activation energy to the face motor processes. Second, they play an important role in regulating face-to-face exchanges with the caretaker. The `drives` play an important role in establishing the `emotional` state of the robot, which is reflected by its facial expression, hence `emotions` play an important role in communicating the state of the robot's "needs" to the caretaker and the urgency for tending to them. It is important that the caretaker find these expressive states compelling as argued in section 2. Certainly, the importance of emotions for believable interactions with artifical systems has already been argued by (Bates, Loyall & Reilly 1992, Cassell 1994, Perlin 1995). Emotions also play an important role in learning during face-to-face exchanges with the caretaker, but we leave the details of this to another paper.

The organization and operation of the `emotion` subsystem is strongly inspired by various theories of emotions in humans (Ekman & Davidson 1994, Izard 1993), and most closely resembles the framework presented in Velasquez (1996), as opposed to the cognitive assessment systems of Ortony, Clore & Collins (1988), Elliot (1992), or Reilly (1996). Kismet has several `emotion` processes. Although they are quite different from emotions in humans, they are designed to be rough analogs — especially with respect to the accompanying facial expressions. As such, each `emotion` is distinct from the others and consists of a family of similar `emotions` which are graded in intensity. For instance, `happiness` can range from being `content` (a baseline activation level) to `ecstatic` (a high activation level). Numerically, the activation level of each `emotion` can range between $[0, max]$ where max is an integer value determined empirically. Although the `emotions` are always active, their intensity must exceed a threshold level before they are expressed externally. When this occurs, the corresponding facial expression reflects the level of activation of the `emotion`. Once an `emotion` rises above its activation threshold, it decays over time back toward the base line level (unless it continues to receive excitatory inputs from other processes or events). Hence, unlike `drives`, `emotions` have an intense expression followed by a fleeing nature. Ongoing events that maintain the activation level slightly above threshold correspond to `moods` in this implementation. For the robot, its drives are a main contributor to establishing its ongoing mood. `Temperaments` are established by setting the gain and bias terms. `Blends` of emotions occur when several compatible emotions are expressed simultaneously. To avoid having conflicting `emotions` active at the same time, mutually inhibitory connections exist between conflicting emotions.

4.3 The Attention System

The attention system acts to direct computational and behavioral resources toward salient stimuli. In an environment suitably complex for interesting learning, perceptual processing will invariably result in many potential target stimuli. In order to determine where to assign resources, the attention system must incorporate raw sensory saliency with motivational influences. Raw sensory saliency cues are equivalent to those "pop-out" effects studied by Triesman (1986), such as color intensity, motion, and orientation for visual stimuli and intensity and pitch for auditory stimuli. The motivational system may bias the selection process, but does not alter the underlying raw saliency of a stimulus (Niedenthal & Kityama 1994). For example, if the robot has become `bored`, it may be more sensitive to visual motion (which may indicate something that would engage the robot) and less sensitive to orientation effects (which are likely to be static background features).

To build a believable creature, the attention system must also implement habituation effects. Infants respond strongly to novel stimuli, but soon habituate and respond less as familiarity increases (Carey & Gelman 1991). This acts both to keep the infant from being continually fascinated with any single object and to force the caretaker to continually engage the infant with slightly new and interesting interactions. For a robot, a habituation mechanism removes the effects of highly salient background objects that are not currently involved in direct interactions as well as placing requirements on the caretaker to maintain interaction with slightly novel stimulation.

4.4 The Behavior System

Borrowing from the behavioral organization theories of Lorenz (1973) and Tinbergen (1951), `drives` within the robot's motivation system cannot satiate themselves. They become satiated whenever the robot is able to evoke the corresponding *consummatory behavior*. For instance, with respect to animals, eating satiates the hunger drive; sleeping satiates the fatigue drive, and so on. At any point in time, the robot is motivated to engage in behaviors that maintain the `drives` within their homeostatic regime. Furthermore, whenever a `drive` moves farther from its desired operation point, the robot becomes more predisposed to engage in behaviors that serve to satiate that `drive` — as the `drive` activation level increases, it passes more of its activation energy to the corresponding consummatory behavior. As long as the consummatory behavior is active, the intensity of the `drive` is reduced toward the homeostatic regime. When this occurs, the `drive` becomes satiated, and the amount of activation energy it passes to the consummatory behavior decreases until the consummatory behavior is eventually released.

For each consummatory behavior, there may also be one or more affiliated *appetitive behaviors*. One can view each appetitive behavior as a separate behavioral strategy for bringing the robot to a state where it can directly activate the desired consummatory behavior. For instance, the case may arise where a given `drive` strongly potentiates its consummatory behavior, but environmental circumstances prevent it from becoming active. In this case, the robot may be able to activate an affiliated appetitive behavior instead, which will eventually enable the consummatory behavior to be activated.

In this implementation, every behavior is modeled as a separate goal-directed process. In general, both internal and external factors are used to compute whether or not they should be activated. The most significant inputs come from the `drive` they act to satiate and from the environment. The activation level of each behavior can range between $[0, max]$ where max is an integer value determined empirically. When a consummatory behavior is active, its output acts to reduce the activation energy of the `drive` it is associated with. When an appetitive behavior is active, it serves to bring the robot into an environmental state suitable for activating the affiliated consummatory behavior.

4.5 The Motor System

The motor system incorporates both motor skills, such as smooth pursuit tracking or saccading, as well as expressive motor acts, such as wiggling the ears or lowering the brow. It commands facial postures to reflect the currently active emotion, and can blend multiple facial postures when several compatible emotions when concurrently active. Each expressive motor act is linked to a corresponding `emotion`. The robot's facial features move analogously to how humans adjust their facial features to express different emotions (Ekman & Friesen 1978), and the robot's ears move analogously to how dogs to move theirs to express motivational state (Milani 1986). The motor system is also responsible for implementing emotional "overlays" over the task based motor skills. This is important for conveying expressiveness through posture – for instance, the robot can look to a given object while conveying apprehension or deliberateness by the way it moves its neck and eye motors as well as its facial motors.

This section has presented a broad overview of the architectural framework of this system. The following sections describe the design details of each of these five systems in greater detail. Specifics of the implementation were chosen to make Kismet an "infant informavoire"[3], that is, to define the robot's nature so that it is driven to learn in a social context. If done properly, the robot will behave in such a way that it can influence the behavior of the caretaker to maintain an interaction the robot can handle, learn from, and use to satisfy its drives.

5 Design of the Perceptual System

Human infants discriminate readily between social stimuli (faces, voices, etc.) and salient non-social stimuli (brightly colored objects, loud noises, large motion, etc.) (Aslin 1987). The perceptual system has been designed to discriminate a subset of both social and non-social stimuli from visual images. As a social stimulus detector, we have implemented a face detector that mimics some of the innate preferences that human infants have for face-like stimuli. We further rely on visual motion detection both to supplement the accuracy of the face detector and as an indicator of the presence of a salient non-social stimulus.

5.1 Perceiving Motion

The motion detection module computes the difference between consecutive wide-angle images within a local field and then uses a region-growing technique to identify contiguous blocks of motion within the difference image. The bounding box of the five largest motion blocks are provided through dual-ported RAM to the motivation system.

The motion detection process receives a digitized 128×128 image from the left wide-angle camera. Incoming images are stored in a ring of three frame buffers; one buffer holds the current image I_0, one buffer holds the previous image I_1, and a third buffer receives new input. The absolute value of the difference between the grayscale values in each image is thresholded to provide a raw motion image ($I_{raw} = \mathcal{T}(|I_0 - I_1|)$). The raw motion image is then filtered with a 3×3 Gaussian function (standard deviation of 2 pixels) in order to filter high-frequency noise.

The filtered image is then segmented into bounding boxes of contiguous motion. The algorithm scans the filtered image, marking all locations which pass threshold with an identifying tag. Locations inherit tags from adjacent locations through a region grow-and-merge procedure (Horn 1986). Once all locations above threshold have been tagged, the tags are sorted based on the number of image pixels that tag marks. The bounding box and centroid of each tagged region is computed, and data on the top five tags are sent to the motivational system.

[3] A term Dan Dennett mentioned to us during conversation.

The motion detection system runs on a single node of the DSP network (labeled “motion”) shown in figure 3. The system operates at 15-30 frames per second, depending on the amount of motion present in the image. Outputs from the motion detection system are displayed on a VGA monitor and sent to the 68332 network through the dual-ported RAM interface.

5.2 Perceiving Faces

The face detection algorithm used here was initially implemented as part of a developmental program for building social skills based on detection of signals of shared attention such as eye direction, pointing gestures, and head position (Scassellati 1998*b*). In that work, our choice of a face detection algorithm was based on three criteria. First, it must be a relatively simple computation that can be performed in real time. Second, the technique must perform well under social conditions, that is, in an unstructured environment where people are most likely to be looking directly at the robot. Third, it should be a biologically plausible technique. Based on these criteria, we selected the ratio template approach described by Sinha (1994). Because these criteria are also applicable to the task specifications for providing perceptual input for social and emotional models discussed in this paper, we elected to use the same algorithm.

The ratio template algorithm was designed to detect frontal views of faces under varying lighting conditions, and is an extension of classical template approaches (Sinha 1996). While other techniques handle rotational invariants more accurately (Sung & Poggio 1994) or provide better accuracy at the cost of greater computation (Turk & Pentland 1991, Rowley, Baluja & Kanade 1995), the simplicity of the ratio template algorithm allows us to operate in real time while detecting faces that are most likely to be engaged in social interactions. Ratio templates also offer multiple levels of biological plausibility; templates can be either hand-coded or learned adaptively from qualitative image invariants (Sinha 1994).

A ratio template is composed of a number of regions and a number of relations, as shown in Figure 5. For each target location in the grayscale peripheral image, a template comparison is performed using a special set of comparison rules. Overlaying the template with a 14 pixel by 16 pixel grayscale image patch at a potential face location, each region is convolved with the grayscale image to give the average grayscale value for that region. Relations are comparisons between region values, for example, between the “left forehead” region and the “left temple” region. The relation is satisfied if the ratio of the first region to the second region exceeds a constant value (in our case, 1.1). This ratio allows us to compare the intensities of regions without relying on the absolute intensity of an area. In Figure 5, each arrow indicates a relation, with the head of the arrow denoting the second region (the denominator of the ratio). This template capitalizes on illumination-invariant observations. For example, the eyes tend to be darker than the surrounding face, and the nose is generally brighter than its surround. We have adapted the ratio template algorithm to process video streams. In doing so, we require the absolute difference between the regions to exceed a noise threshold, in order to eliminate false positive responses for small, noisy grayscale values. Figure 6 shows a sample image processed by the face detection algorithm.

The ratio template algorithm can easily be modified to detect faces at multiple scales. Multiple nodes of the parallel network run the same algorithm on different sized input images, but without changing the size of the template. This allows the system to respond more quickly to faces that are closer to the robot, since closer faces are detected in smaller images which require less computation. With this hardware platform, a 64×64 image and a 14×16 template can be used to detect faces within approximately three to six feet of the robot. The same size template can be used on a 128×128 image to find faces within approximately twelve feet of the robot.

5.2.1 Improving the Speed of Face Detection

To improve the speed of the ratio template algorithm, we have implemented two optimizations: an early-abort scheme and a motion-based prefilter.

At the suggestion of Sinha (1997), we further classified the relations of our ratio-template into two categories: eleven essential relations, shown as solid arrows in Figure 5, and twelve confirming relations, shown as dashed arrows. We performed a post-hoc analysis of this division upon approximately ten minutes of video feed in which one of three subjects was always in view. For this post-hoc analysis, an arbitrary threshold of eighteen of the twenty-three relations was required to be classified as a face. This threshold eliminated virtually all false positive detections while retaining at least one detected face in each image. An analysis of the detected faces indicated that at least ten of the eleven essential relations were always satisfied. None of the confirming relations achieved that level of specificity. Based on this analysis, we established a new set of thresholds for face detection: ten of the eleven essential relations and eight of the twelve confirming relations must be satisfied. As soon as two or more of the essential relations have failed, we can reject the location as a face. This increases the speed of our computation by a factor of 4, as shown in Table 1, without any observable decrease in performance.

To further increase the speed of our computation, we use a pre-filtering technique based on motion. The pre-filter allows us to search only locations that are likely to contain a face. Consecutive images are differenced, thresholded, and then convolved with a 14×16 kernel of unitary value (the same size as the ratio template) in order to generate the average amount of super-threshold movement for each potential face location. If that average motion value for a location exceeds threshold, then that location is a candidate for face detection. For each incoming frame, a location is a potential target for the face detection routine if it has had motion within the last five frames, if the ratio template routine verified a face in that location within the last five frames, or if that location had not been checked for faces within the last three seconds. In this way, we capture faces that have just entered the field of view (through the motion clause) and faces that have stopped moving (through the past history clause). The prefilter also resets every three seconds, allowing the system to re-acquire faces that have dropped below the noise threshold. The prefilter automatically resets any time the active system moves, since this generates induced motion of the visual field. This filtering technique increased the speed of the face detection routines by a factor of five for 64×64 images and a factor of eight for 128×128 images (see Table 1). The smaller image size appeared to saturate at 20 Hz due to constant computational loads in the rest of the system, primarily from drawing display images to a VGA display. The filtering technique greatly reduced the number of background locations to be searched without any observable loss of accuracy.

5.2.2 Evaluation of Ratio Templates

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

To test lighting invariance, we considered only the images with an upright head position at a single scale, giving a test set of 48 images under lighting conditions with the primary light source at 90 degrees, 45 degrees, and head-on. Figure 7 shows the images from two of the subjects under each lighting condition. The ratio template algorithm detected 34 of the 48 test faces. Of the 14 faces that were missed, nine were the result of three subjects that failed to be detected under any lighting conditions. One of these subjects had a full beard, while another had very dark rimmed glasses, both of which seem to be handled poorly by the static detection algorithm. Of the remaining five misses, two were from the 90 degree lighting condition, two from the 45 degree lighting condition, and one from the head-on condition. While this detection rate (71%) is considerably lower than other face detection schemes (Rowley et al. 1995, Turk & Pentland 1991, Sung & Poggio 1994), this result is a poor indicator of the performance of the algorithm in a complete, behaving system, as we will see below.

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

Quantitative analysis of behaving systems difficult, and often misleading (Brooks 1991). Our system does not require a completely general-purpose face recognition engine. In a real-world environment, the caretaker is generally cooperative. She is attempting to be seen by the robot, keeping her attention focused on the robot, facing toward it, and often unconsciously moving to try to attract its attention. Further, the system need not be completely accurate on every timestep; its behavior need only converge to the correct solution. If the system can adequately recognize these situations, then it has fulfilled its purpose.

While this algorithm performed relatively poorly on a standard test set of static face images, this measurement was a poor indicator of how the algorithm would perform on live video streams. By utilizing a pair of learned sensorimotor mappings, our system was capable of saccading to faces and extracting high resolution images of the eye on 94% of trials (Scassellati 1998*b*). Figure 8 shows six of the faces detected from that set of trials. However, even this statistic was misleading, since the behavior of the overall system eventually corrected for trials where the first saccade missed the target. To further evaluate behaving systems in complex environments, more refined observation techniques are necessary. However, for the purposes of this paper, the face detection algorithm has been more than adequate.

6 Design of the Motivation System

The robot's motivational system is composed of two inter-related subsystems. One subsystem implements the robot's `drives`, another implements its `emotions` and `expressive states`. Figure 9 shows the current system implementation for the entire behavior engine.

6.1 The Drive Subsystem

For an animal, adequately satisfying its drives is paramount to survival. Similarly, for the robot, maintaining all its `drives` within their homeostatic regime is a never-ending, all important process.

Currently, the robot has three basic `drives`:

- `Social drive`: One `drive` is to be social, that is, to be in the presence of people and to be stimulated by people. This is important for biasing the robot to learn in a social context. On the under-whelmed extreme the robot is `lonely`; it is predisposed to act in ways to establish face-to-face contact with people. If left unsatiated, this `drive` will continue to intensify toward the `lonely` end of the spectrum. On the over-whelmed extreme, the robot is `asocial`; it is predisposed to act in ways to avoid face-to-face contact. The robot tends toward the `asocial` end of the spectrum when a person is over-stimulating the robot. This may occur when a person is moving too much or is too close to the camera.

- `Stimulation drive`: Another `drive` is to be stimulated, where the stimulation can either be generated externally by the environment or internally through spontaneous self-play. On the under-whelmed end of this spectrum, the creature is `bored`. This occurs if the creature has been inactive or unstimulated over a period of time. On the over-whelmed part of the spectrum, the creature is `confused`. This occurs when the robot receives more stimulation than it can effectively assimilate, and predisposes the robot to reduce its interaction with the environment, perhaps by closing its eyes or turning its head away from the stimulus. In the future, this `drive` will also be relevant for learning; this `drive` will tend toward the `bored` end of the spectrum if the current interaction becomes very predictable for the robot. This will bias the robot to engage in new kinds of activities and encourage the caretaker to challenge the robot with new interactions.

- `Fatigue drive`. This `drive` is unlike the others in that its purpose is to allow the robot to shut out the external world instead of trying to regulate its interaction with it. While the creature is "awake", it receives repeated stimulation from the environment. As time passes this `drive` approaches the `exhausted` end of the spectrum. Once the intensity level exceeds a certain threshold, it is time for the robot to "sleep". In the future, this will be the time for the robot to consolidate its learned anticipatory models and integrate them with the rest of the internal control structure. While the robot "sleeps", the `drive` returns to the homeostatic regime, and the robot awakens.

6.2 The Emotion and Expressive States Subsystem

So far, there are a total of eight `emotions` and `expressive states` implemented in this system, each as a separate transducer process. The overall framework of the emotion system shares strong commonality with that of Velasquez (1996), although its function is specifically targeted for social exchanges and learning. The robot has analogs of five primary emotions in humans: `anger`, `disgust`, `fear`, `happiness`, and `sadness`. The robot also has three `expressive states` that do not correspond to human emotions, but do play an important role in human learning and social interaction: `surprise`, `interest`, `excitement`. Many experiments in developmental psychology have shown that infants show surprise when witnessing an unexpected or novel outcome to a familiar event (Carey & Gelman 1991). Furthermore, parents use their infant's display of excitement or interest as cues to regulate their interaction with them (Wood et al. 1976).

In humans, four factors serve to elicit emotions: neurochemical, sensorimotor, motivational, and cognitive factors (Izard 1993). In this system, emphasis has been placed on how `drives` and other `emotions` contribute to a given `emotion's` level of activation.

- `Drives`: Recall that each `drive` is partitioned into three regimes: homeostatic, overwhelmed or under-whelmed. For a given drive, each region potentiates a different emotion and hence a different facial expression. In this way the facial expressions provide cues as to what drive is out of balance and how the caretaker should respond to correct for it.

- Other `emotions`: The influence from other `emotions` serve to prevent conflicting `emotions` from becoming active at the same time. To implement this, conflicting `emotions` have mutually inhibitory connections between them. For instance, inhibitory connections exist between `happiness` and `sadness`, between `disgust` and `happiness`, and between `happiness` and `anger`.

In general, when a `drive` is in its homeostatic regime, it potentiates positive `emotions` such as `happiness` or `interest`. The accompanying expression tells the caretaker that the interaction is going well and the robot is poised to play and learn. When a `drive` is not within the homeostatic regime, negative `emotions` are potentiated (such as `anger`, `disgust`, or `sadness`) which produces signs of distress on the robot's face. The particular sign of distress provides the caretaker with additional cues as to what is "wrong" and how she might correct for it. For example, overwhelming social stimuli (such as a rapidly moving face) produce signs of `disgust` – an asocial response. In contrast, overwhelming non-social stimuli (such as a rapidly moving ball) produce signs of `fear`. Infants often show signs of anxiety when placed in a confusing environment.

Note that the same sort of interaction can have a very different "emotional" effect on the robot depending on the drive context. For instance, playing with the robot while all `drives` are within the homeostatic regime elicits `happiness`. This tells the caretaker that playing with the robot is a good interaction to be having at this time. However, if the `fatigue drive` is deep into the `exhausted` end of the spectrum, then playing with the robot actually prevents the robot from going to sleep. As a result, the `fatigue drive` continues to increase in intensity. When high enough, the `fatigue drive` begins to potentiate `anger`. The caretaker may interpret this as the robot acting "cranky" because it is "tired". In the extreme case, `fatigue` may potentiate `anger` so strongly that the robot displays "fury". The caretaker may construe this as the robot throwing a "tantrum". Normally, the caretaker would back off before this point and allow the `sleep` behavior to be activated.

Important near-term extensions to this subsystem include adding a variety of sensorimotor elicitors so the robot can respond emotionally to various perceptual stimuli. For instance, the robot should show immediate displeasure to very intense stimuli, show interest to particularly salient stimuli, and show surprise to strong and suddenly appearing stimuli.

7 Design of the Attention System

The current implementation has a very simplistic attentional mechanism. To limit the computational requirements, the robot processes only the most salient face stimulus, which is the target location that gives the best quantitative match to the ratio template, and to the five most salient motion stimuli, which are the five largest contiguous regions of motion. All other output from these perceptual processes are suppressed. Note that this attentional process does not currently limit the computational requirements of perception, nor does it account for habituation effects or for influences from the motivational system. However, this simplistic system does limit the computation necessary for behavior selection. A more complex attention system that incorporates habituation, influences from the motivational system, and additional sensory inputs is currently under construction.

8 Design of the Behavior System

For each `drive` there is an accompanying consummatory behavior. Ideally, it becomes active when the `drive` enters the under-whelmed regime and remains active until it returns to the homeostatic regime. The consummatory behaviors are as follows:

- `Socialize` acts to move the `social drive` toward the `asocial` end of the spectrum. It is potentiated more strongly as the `social` drive approaches the `lonely` end of the spectrum. Its activation level increases above threshold when the robot can engage in social interaction with a person, that is, when it can obtain a face stimulus at a reasonable activation level. The behavior remains active for as long as this interaction is maintained. Only when active does it act to reduce the intensity of the drive. When the interaction is of suitable intensity, the drive approaches the homeostatic regime and remains there. When the interaction is too intense, the drive will pass the homeostatic regime and move into the `asocial` regime.

- `Play` acts to move the `stimulation drive` toward the `confused` end of the spectrum. It is potentiated more strongly as the `stimulation drive` approaches the `bored` end of the spectrum. The activation level increases above threshold when the robot can engage in some sort of stimulating interaction, in this case, by observing a non-face object that moves gently. It remains active for as long as the robot maintains the interaction. While active it continues to move the `drive` toward the `confused` end of the spectrum if the interaction is too intense. If the interaction is appropriate, the `drive` will remain in the homeostatic regime.

- `Sleep` acts to satiate the `fatigue drive`. When the `fatigue drive` reaches a specified level, the `sleep` consummatory behavior turns on and remains active until the `fatigue drive` is restored to the homeostatic regime. When this occurs, it is released and the robot “wakes up”.

`Sleep` also serves a special “motivation reboot” function for the robot. When active, it not only restores the `fatigue drive` to the homeostatic regime, but all the other drives as well. If any `drive` moves far from its homeostatic regime, the robot displays stronger and stronger signs of distress, which eventually culminates in extreme `anger` if left uncorrected. This expressive display is a strong sign to the caretaker to intervene and help the robot. If the caretaker fails to act appropriately and the drive reaches an extreme, a protective mechanism activates and the robot eliminates external stimulation by going to `sleep`. This extreme self-regulation method allows the robot to restore all its `drives` by itself. A similar behavior is observed in infants; when they are in extreme distress, they may fall into a disturbed sleep (Bullowa 1979).

In the simplest case, each drive and its satiating consummatory behavior are connected as shown in figures 10, 11, and 12. Both the `drive` and the consummatory behavior are modeled as transducers where the output is simply the current activation energy. As shown, the output of a `drive` is fed into an excitatory input of its consummatory behavior. Hence, as the `drive` grows in intensity, it potentiates the activation level of the consummatory behavior more and more. When the activation level rises above threshold, the consummatory behavior is active and is expressed through the robot's behavior. As the robot performs this behavior, the output of the consummatory behavior is fed back into an inhibitory input of the `drive`. This acts to reduce the `drive`'s intensity level. As the `drive`'s intensity decreases, it potentiates the consummatory behavior less and less. Finally, when the `drive` is restored to the homeostatic regime, the activation level of the consummatory behavior falls below its activation threshold and it is deactivated.

Two of the three consummatory behaviors cannot be activated by the intensity of the `drive` alone. Instead, they require a special sort of environmental interaction to become active. For instance, `socialize` cannot become active without the participation of a person. (Analogous cases hold for `play`.) Furthermore, it is possible for these behaviors to become active by the environment alone if the interaction is strong enough. This has an important consequence for regulating the intensity of interaction. For example, if the nature of the interaction is too intense, the `drive` may move into the over-whelmed regime. In this case, the `drive` is no longer potentiating the consummatory behavior; the environmental input alone is strong enough to keep it active. When the `drive` enters the over-whelmed regime, the system is strongly motivated to engage in behaviors that act to stop the stimulation. For instance, if the caretaker is interacting with the robot too intensely, the `social drive` may move into the `asocial` regime. When this occurs, the robot displays an expression of displeasure, which is a cue for the caretaker to stop.

9 Design of the Motor System

Our current system design has incorporated expressive motor actions for each `emotion`. Additionally, we have implemented the hardware control for various motor skills, such as smooth pursuit tracking and saccadic eye movement (Scassellati 1998*a*), but have yet to incorporate these skills into the behavior engine.

As described in section 3, the robot currently has eleven face actuators that move two eyebrows and two ears, each with two degrees of freedom, as well as two eyelids and a mouth, each with one degree of freedom. Each eyebrow can be raised or lowered, and can arc in toward the nose or out toward the ear. The ears can be raised or lowered and rotate forward or backward. The eyelids and lower jaw can be raised or lowered.

The low-level face motor primitives are separate transducer processes that control the position and velocity of each degree of freedom. One level above these processes exist coordinated motion processes which control of coordinated movements of the facial feature such as wiggling the ears or eyebrows independently, arching both brows inward, raising the brows, and so forth. Generally, they are coordinated motions used in common facial expressions. Above these coordinated motion processes are the face expression processes. These direct all facial features to show a particular expression. For each expression, the facial features move to a characteristic configuration, with the intensity depending on the intensity of the emotion evoking the expression. In general, the more intense the expression, the facial features move more quickly to more extreme positions. Blended expressions are computed by taking a weighted average of the facial configurations corresponding to each evoked emotion. In general, expressive acts may modify the task based motor skills (such as looking at a particular object) or overall postures (eye and neck position) to convey different emotional states. This has yet to be implemented.

10 Experiments and Results

A series of experiments was performed with the robot using the behavior engine shown in figure 9. The total system consists of three `drives` (`fatigue`, `social`, and `stimulation`), three consummatory behaviors (`sleep`, `socialize`, and `play`), two visually-based percepts ("face" and "non-face"), five `emotions` (`anger`, `disgust`, `fear`, `happiness`, `sadness`), two expressive states (`tiredness` and `interest`), and their corresponding facial expressions. More detailed schematics for the "stimulation" circuit, the "social" circuit, and the "fatigue" circuit are shown in figures 10, 11, and 12 respectively.

Each experiment involved a human interacting with the robot either through direct face-to-face interaction, by waving a hand at the robot, or using a toy to play with the robot. The toys are shown in figure 1; one is a small plush black and white cow and the other is an orange plastic slinky. The perceptual system classifies these interactions into two classes: *face stimuli* and *non-face stimuli*. The face detection routine classifies both the human face and the face of the plush cow as face stimuli, while the waving hand and the slinky are classified as non-face stimuli. Additionally, the motion generated by the object gives a rating of the stimulus intensity. The robot's facial expressions reflect its ongoing motivational state (i.e. it's mood) and provides the human with visual cues as to how to modify the interaction to keep the robot's `drives` within homeostatic ranges.

In general, as long as all the robot's `drives` remain within their homeostatic ranges, the robot displays "interest". This cues the human that the interaction is of appropriate intensity. If the human engages the robot in face-to-face contact while its `drives` are within their homeostatic regime, the robot displays "happiness". However, once any `drive` leaves its homeostatic range, the robot's "interest" and/or "happiness" wane(s) as it grows increasingly distressed. As this occurs, the robot's expression reflects its distressed state. This visual cue tells the human that all is not well with the robot, whether the human should switch the type of stimulus, and whether the intensity of interaction should be intensified, diminished or maintained at its current level.

For all of these experiments, data was recorded on-line in real-time during interactions between a human and the robot. Figures 13 through 18 plot the activation levels of the appropriate `emotions`, `drives`, behaviors, and percepts. `Emotions` are always plotted together with activation levels ranging from 0 to 2000. Percepts, behaviors, and `drives` are often plotted together. Percepts and behaviors have activation levels that also range from 0 to 2000, with higher values indicating stronger stimuli or higher potentiation respectively. `Drives` have activations ranging from -2000 (the over-whelmed extreme) to 2000 (the under-whelmed extreme).

10.1 Non-face stimuli experiments

Figures 13 and 14 illustrate the influence of the `stimulation drive` on the robot's motivational and behavioral state when interacting with a salient non-face stimulus. The activation level of the robot's `play` behavior cannot exceed the activation threshold unless the human interacts with the robot with sufficient intensity – low intensity interaction will not trigger the `play` behavior even if highly potentiated by the `stimulation drive`. If the interaction is intense, even too intense, the robot's `play` behavior remains active until the human either stops the activity, or the robot takes action to end it.

For the waving hand experiment, a lack of interaction before the start of the run ($t \leq 0$) places the robot in a "sad" emotional state as the `stimulation drive` lies in the `bored` end of the spectrum for activations ≥ 400. From $5 \geq t \geq 25$ a waving hand stimulates the robot within the acceptable intensity range ($400 \geq$ $stimulus$ $\geq$1600) on average. This corresponds to giving the robot small, gentle waves. This amount of stimulus causes the `stimulation drive` to diminish until it resides within the homeostatic range, and a look of "interest" appears on the robot's face. From $25 \geq t \geq 45$ the stimulus maintains a desirable intensity level, the `drive` remains in the homeostatic regime, and the robot maintains "interest". At $45 \geq t \geq 70$ the hand stimulus intensifies to large, sweeping motions which overwhelm the robot ($intensity \geq 1600$). This causes the `stimulation drive` to migrate toward the over-whelmed end of the spectrum. As the `drive` approaches the over-whelmed extreme, the robot's face displays an intensifying expression of "fear". Around $t = 75$ the robot looks "terrified" at an emotional level of 1500. The experimenter responds by stopping the waving stimulus until the robot "calms" down a bit as exhibited by a lessening of its "fear" expression, and then resumes the stimulation within the acceptable range. Consequently, the `stimulation drive` returns to the homeostatic regime and the robot displays "interest" again. At $t \geq 105$ the waving stimulus stops for the remainder of the run. Because the robot is under-stimulated the `stimulation drive` moves into the `bored` end of the spectrum and an expression of "sadness" reappears on the robot's face.

The slinky experiment was conducted in a similar fashion. As in the previous case, the robot is placed into a `bored` state before the experiment begins. At $t = 5$ the robot is shown small slinky motions which correspond to an acceptable intensity. Occasionally the slinky motion is too intense ($t = 30$ and $t = 35$), but on average the motion is acceptable. As a result, the `stimulation drive` is restored to the homeostatic regime and the robot looks "interested". At $75 \geq t \geq 105$ the experimenter moves the slinky in large sweeping motions which are too vigorous for the robot. Consequently the `drive` moves deep into the over-whelmed regime. When the `drive` intensity passes -1600, an expression of "anger" is blended with the intensifying look of "fear". At $t = 105$, the experimenter stops the slinky motion completely and allows the robot to "calm" down a bit, and then resumes small slinky motions. In response, the `drive` returns to the homeostatic regime and the robot appears "interested" again. At $t \geq 150$ the slinky motion ceases, and this lack of stimulation causes the `drive` to move back into the under-whelmed end of the spectrum, and an expression of "sadness" returns to the robot's face.

10.2 Face stimuli experiments

Figures 15 and 16 illustrate the influence of the `social drive` on the robot's motivational and behavioral state when interacting with a face stimulus. The robot's `socialize` behavior cannot become active unless a human interacts with the robot with sufficient intensity – low intensity interaction will not trigger the `socialize` behavior even if highly potentiated by the `social drive`. Whenever the interaction exceeds this base threshold ($intensity \geq 400$), the robot's `socialize` behavior remains active until either the human or the robot terminates the interaction.

Figure 15 shows the interaction of the robot with a human face stimulus. Before the run begins, the robot is not shown any faces so that the `social drive` lies in the `lonely` regime and the robot displays an expression of "sadness". At $t = 10$ the experimenter makes face-to-face contact with the robot. From $10 \geq t \geq 58$ the face stimulus is within the desired intensity range. This corresponds to small head motions, much like those made when engaging a person in conversation. As a result, the `social drive` moves to the homeostatic regime, and a look of "interest" and "happiness" appear on the robot's face. From $60 \geq t \geq 90$ the experimenter begins to sway back and forth in front of the robot. This corresponds to a face stimulus of over-whelming intensity, which forces the `social drive` into the `asocial` regime. As the drive intensifies toward a value of -1800, first a look of "disgust" appears on the robot's face, which grows in intensity and is eventually blended with "anger". From $90 \geq t \geq 115$ the experimenter turns her face away so that it is not detected by the robot. This allows the `drive` to recover back to the homeostatic regime and a look of "interest" returns to the robot's face. From $115 \geq t \geq 135$ the experimenter re-engages the robot in face-to-face interaction of acceptable intensity and the robot, and the robot responds with an expression of "happiness". From $135 \geq t \geq 170$ the experimenter turns away from the robot, which causes the `drive` to return to the `lonely` regime and redisplay "sadness". For $t \geq 170$ the experimenter re-engages the robot in face-to-face contact, which leaves the robot in an "interested" and "happy" state at the conclusion of the run.

Figure 16 shows the interaction of the robot with the plush cow toy. Because the face detector picks out the cow's face, the cow is treated as a social stimulus and thereby affects the `social drive`. This experimental run follows the same format as that for the human face stimulus. The run begins with the `social drive` within the `lonely` regime and the robot looking "sad". At $t = 5$ the experimenter shows the robot the cow's face and moves the cow in small gentle motions. This corresponds to a stimulus of acceptable intensity level which restores the `drive` to the homeostatic regime. As a result the robot appears "interested" and "happy". From $50 \geq t \geq 78$ the experimenter begins swinging the cow quickly in front of the robot's face. Because the stimulus is too intense, the `drive` moves into the `asocial` regime and the robot expression of "disgust" intensifies until eventually blended with "anger" as well. At $t = 78$ the experimenter removes the cow from the robot's visual field and allows the `drive` to return to the homeostatic regime. From $98 \geq t \geq 118$ the cow's face is shown to the robot again which maintains the `drive` within the homeostatic regime and the robot displays "interest" and "happiness". From $118 \geq t \geq 145$ the cow's backside is shown to the robot. The lack of a face stimulus causes the `social drive` to return to the `lonely` regime, but at $t \geq 145$ the cow is turned to face the robot and the `drive` is restored to the homeostatic regime until the conclusion of the run. The run ends with the robot in a "happy" and "interested" state.

10.3 Sleep and over-stimulation experiments

As discussed in previous sections, infants fall into a disturbed sleep when put into an extremely anxious state for a prolonged time. Analogously for the robot, if the interaction is over-whelming for long periods of time, the robot will first show increasing signs of "disgust", eventually blending with increasingly intense signs of "anger", as the social drive continues to move toward the over-whelmed end of the spectrum. Figure 17 shows one example of this effect. When no relief is encountered and the drive hits its outer limit ($t = 30$), the robot goes into an emergency sleep mode. As discussed previously, sleeping serves as a sort of "motivational reboot" for the robot by restoring all drives to their homeostatic ranges. Hence, upon "awakening", the robot is in a balanced, "interested" state.

Figure 18 illustrates the influence of the fatigue drive on the robot's motivational and behavioral state when interacting with a human. Over time, the fatigue drive increases toward the exhausted end of the spectrum. As the robot's level of "fatigue" increases, the robot displays stronger signs of being "tired". At time step $t = 95$, the fatigue drive moves above the threshold value of 1600 which is sufficient to activate the sleep behavior when no other interactions are occurring. The robot remains "asleep" until all drives are restored to their homeostatic ranges. Once this occurs, the activation level of the "sleep" behavior decays until the behavior is no longer active and the robot "wakes up" in an "interested" state. However, at time step $t = 215$, the plot shows what happens if a human continues to interact with the robot despite its "fatigued" state. The robot cannot fall asleep as long as a person interacts with it because the play behavior remains active (note the mutually inhibitory connections in figure 12). If the fatigue drive exceeds threshold and the robot cannot fall "asleep", the robot begins to show signs of "anger". Eventually the robot's level of "anger" reaches an intense level of 1800, and the robot appears rageful – akin to throwing a "tantrum". Still the human persists with the interaction, but eventually the robot's fatigue level reaches near maximum and emergency actions are taken by the robot to force an end to the interaction. The robot falls into a distressed sleep to restore its drives.

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

11 Summary

We have presented a framework (heavily inspired from work in ethology, psychology, and cognitive development) for designing behavior engines for autonomous robots specifically geared to regulate human-robot interaction. We have shown how the perceptions, drives, emotions, behaviors, and facial expressions influence each other to establish and maintain social interactions that can provide suitable learning episodes, i.e., where the robot is proficient yet slightly challenged, and where the robot is neither under-stimulated nor over-stimulated by its interaction with the human. With a specific implementation, we demonstrated how the system engages in a mutually regulatory interaction with a human while distinguishing between stimuli that can be influenced socially (faces) and those that cannot (motion).

The specifics of learning in a social context (what is learned and how it is learned) were not addressed in this paper. That is the subject of future work, which will include tuning and adjusting this early motivation system to appropriately regulate the intensity of interaction to benefit the learning process. Additional areas of future investigation include the implementation of a selective attentional mechanism, additional motor skills such as smooth pursuit tracking and saccadic eye movement, vocalization capabilities, and additional perceptual capabilities including detecting facial gestures, emotive cues of the caretaker from visual and auditory data streams, or attentional markers such as eye direction and pointing gestures. As such, we are continuing to lay the foundation upon which the learning of early communication skills (turn taking, shared attention, vocalizations having shared meaning) can take place.

12 Acknowledgments

Support for this research was provided by a MURI grant under the Office of Naval Research contract N00014–95–1–0600. The motivational system was designed and largely implemented during a visiting appointment at the Santa Fe Institute.

References

Arkin, R. (1988), Homeostatic control for a mobile robot: dynamic replanning in hazardous environments, *in* W. Wolfe, ed., 'Mobile Robots III', SPIE–The International Society for Optical Engineering, Bellingham, WA, pp. 407–413.

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

Balch, R. & Arkin, R. (1994), 'Communication in Reactive Multiagent Robotic Systems', *Autonomous Robots* pp. 27–52.

Bates, J., Loyall, B. & Reilly, S. (1992), An architecture for action, emotion, and social behavior, Technical Report CMU-CS-92-144, CMU, Pittsburgh, PA.

Billard, A. & Dautenhahn, K. (1997), Grounding Communication in Situated, Social Robots, Technical Report UMCS-97-9-1, University of Manchester.

Blumberg, B. (1996), Old Tricks, New Dogs: Ethology and Interactive Creatures, PhD thesis, MIT.

Breazeal(Ferrell), C. (1998), A Motivational System for Regulating Human-Robot Interaction, *in* 'Proceedings of AAAI98'.

Brooks, R. (1986), 'A robust layered control system for a mobile robot', *IEEE Journal of Robotics and Automation* **RA-2**, 253–262.

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

Brooks, R. A., Ferrell, C., Irie, R., Kemp, C. C., Marjanovic, M., Scassellati, B. & Williamson, M. (1998), Alternative Essences of Intelligence, *in* 'Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)', AAAI Press.

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

Carey, S. & Gelman, R. (1991), *The Epigenesis of Mind*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Cassell, J. (1994), Animated Conversation, *in* 'Proceedings of SIGGRAPH 94'.

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

Coombs, D. J. (1992), Real-Time Gaze Holding in Binocular Robot Vision, Technical Report TR415, U. Rochester.

Ekman, P. & Davidson, R. (1994), *The Nature of Emotion: Fundamental Questions*, Oxford University Press, New York.

Ekman, P. & Friesen, W. (1978), *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, CA.

Elliot, C. D. (1992), The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System, PhD thesis, Institute for the Learning Sciences.

Goldstein, E. B. (1989), *Sensation and Perception*, Wadsworth Publishing Company.

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

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

Izard, C. (1993), Four Systems for Emotion Activation: Cognitive and Noncognitive Processes, *in* 'Psychological Review', Vol. 100, pp. 68–90.

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

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

Maes, P. (1990), 'Learning Behavior Networks from Experience', *ECAL90*.

Mataric, M. (1995), 'Issues and approaches in the design of collective autonomous agents', *Robotics and Autonomous Systems* **16**(2–4), 321–331.

McFarland, D. & Bosser, T. (1993), *Intelligent Behavior in Animals and Robots*, MIT Press, Cambridge, MA.

Milani, M. (1986), *The Body Language and Emotion of Dogs*, William Morrow and Company, New York, NY.

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

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

Niedenthal, P. & Kityama, S. (1994), *The Heart's Eye: Emotional influences in Perception and Attention*, Academic Press.

Ortony, A., Clore, G. & Collins, A. (1988), *The Cognitive Structure of Emotion*, Cambridge University Press.

Perlin, K. (1995), 'Real Time Responsive Animation with Personality', *IEEE Transactions on Visualization and Computer Graphics*.

Reilly, S. (1996), Believable Social and Emotional Agents, PhD thesis, CMU School of Computer Science, Pittsburgh, PA.

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

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

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

Scassellati, B. (1998*b*), Finding Eyes and Faces with a Foveated Vision System, *in* 'Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)', AAAI Press.

Scassellati, B. (1998*c*), Imitation and Mechanisms of Shared Attention: A Developmental Structure for Building Social Skills, Technical Report Technical Report 98-1-005, University of Aizu, Aizu-Wakamatsu, Japan.

Sharkey, P. M., Murray, D. W., Vandevelde, S., Reid, I. D. & McLauchlan, P. F. (1993), 'A modular head/eye platform for real-time reactive vision', *Mechatronics Journal* **3**(4), 517–535.

Sinha, P. (1994), 'Object Recognition via Image Invariants: A Case Study', *Investigative Ophthalmology and Visual Science* **35**, 1735–1740.

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

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

Steels, L. (1995), 'When are robots intelligent autonomous agents', *Robotics and Autonomous Systems* **15**(1–2), 3–9.

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

Tinbergen, N. (1951), *The Study of Instinct*, Oxford University Press, New York.

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.

Triesman, A. (1986), 'Features and Objects in Visual Processing', *Scientific American* **255**, 114B–125.

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

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

Velasquez, J. (1996), Cathexis, A Computational Model for the Generation of Emotions and their Influence in the Behavior of Autonomous Agents, Master's thesis, MIT.

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

Image Size	Detection Method		
	Template	+ Early-Reject	+ Prefilter
64×64	1 Hz	4 Hz	20 Hz
128×128	.25 Hz	1 Hz	8 Hz

Table 1: Processing speed for two image sizes with various optimizations. The original ratio template method is enhanced by a factor of four with the addition of the early-reject optimization, and by an additional factor of five to eight by the prefilter optimization. The system saturated near 20 Hz due to constant computational loads in other parts of the network. All statistics are for a single TMS320C40 node with no other processes.

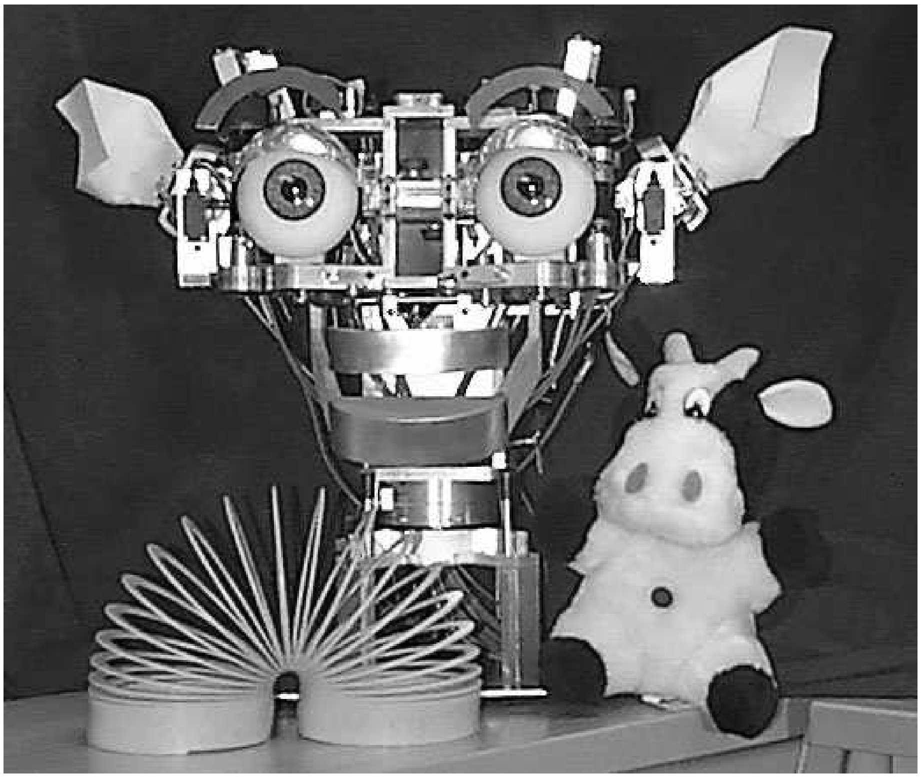

Figure 1: Kismet with toys. Kismet has an active stereo vision system with color CCD cameras mounted inside the eyeballs. There are also a variety of facial features which give the robot its expressive capabilities.

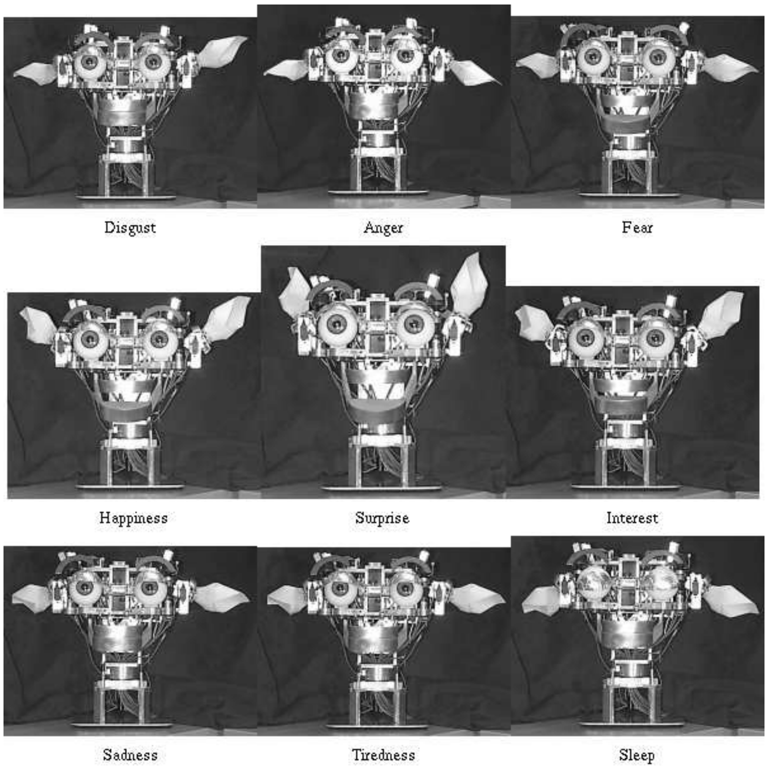

Figure 2: Static extremes of Kismet's facial expressions. During operation, the 11 degrees-of-freedom for the ears, eyebrows, mouth, and eyelids vary continuously with the current emotional state of the robot.

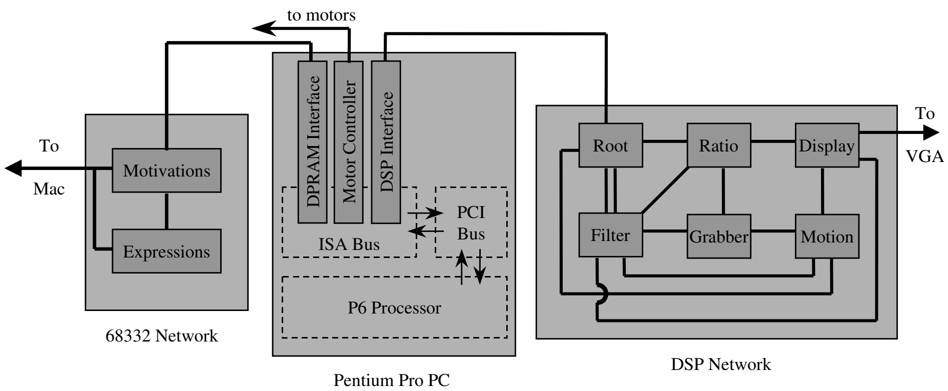

Figure 3: Computational hardware utilized by Kismet. A network of digital signal processors acts as the sensory processing engine and implements the perception system, the attention system, and part of the motor system. This network is attached to two 68332-based microcontrollers that implement the motivational, behavioral, and remainder of the motor systems.

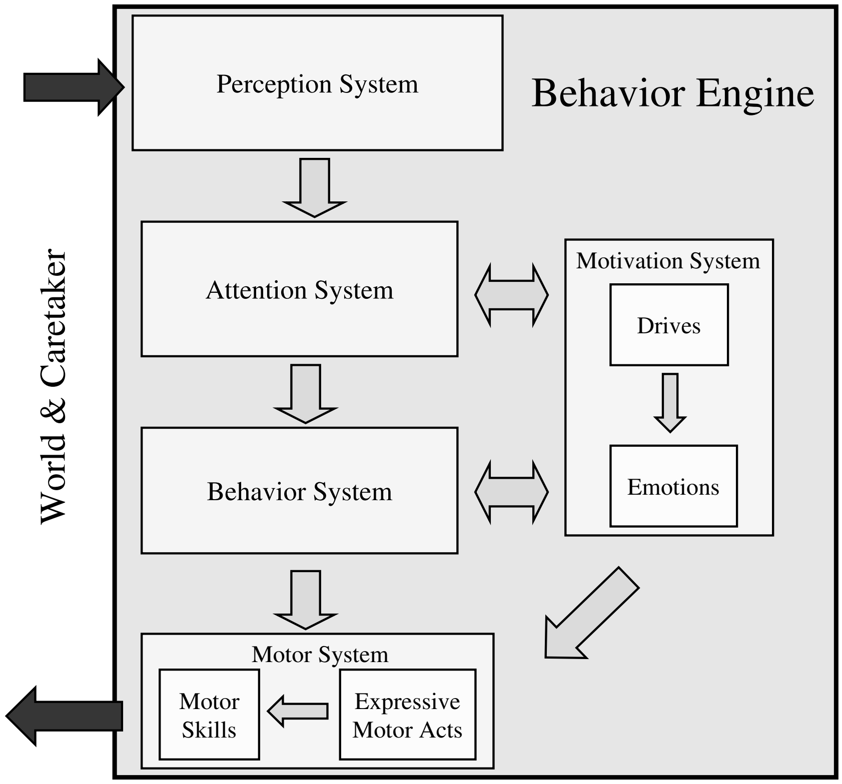

Figure 4: A framework for designing behavior engines. Five systems interact to enable the robot to behave coherently. The perception system extracts salient features from the world, the motivation system maintains internal state in the form of `drives` and `emotions`, the attention system determines saliency based upon perception and motivation, the behavior system selects a set of coherent actions, and the motor system realizes these behaviors as facial expressions and other motor skills.

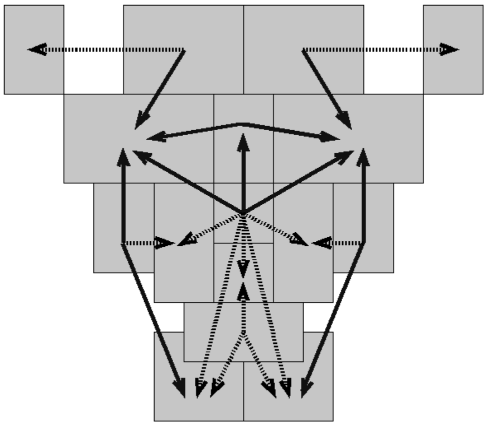

Figure 5: A 14 pixel by 16 pixel ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows). Essential relations are shown as solid arrows while confirming relations are shown as dashed arrows. Adapted from Sinha (1996).

Figure 6: An example face in a cluttered environment. The 128x128 grayscale image was captured by the active vision system, and then processed by the pre-filtering and ratio template detection routines. One face was found within the image, and is shown outlined.

Figure 7: Six of the static test images from Turk and Pentland (1991) used to evaluate the ratio template face detector. Each face appears in the test set with three lighting conditions, head-on (top), from 45 degrees (middle), and from 90 degrees (bottom). The ratio template correctly detected 71% of the faces in the database, including each of these faces except for the middle image from the first column. However, this was a poor indicator of overall performance (see text).

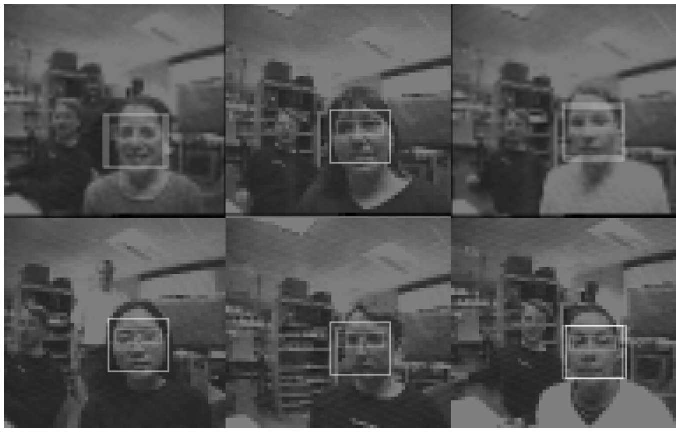

Figure 8: Six detected faces. Only faces of a single scale (roughly within four feet of the robot) are shown here.

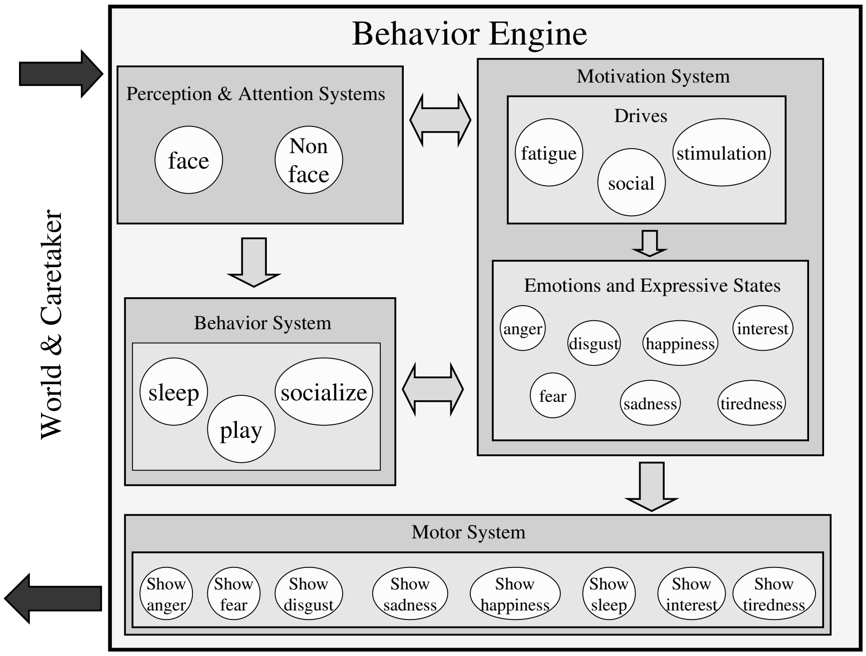

Figure 9: Implementation of the behavior engine framework used in the experiments presented here. There are two percepts, resulting from face-like stimuli and non-face stimuli. The motivation system contains three `drives` (`fatigue`, `social`, and `stimulation`) and eight `emotions` and expressive states (`anger`, `disgust`, `happiness`, `interest`, `fear`, `sadness`, and `tiredness`) each of which can be expressed through the motor system. These percepts and motivations influence the selection of the three behaviors (`sleep`, `play`, and `socialize`).

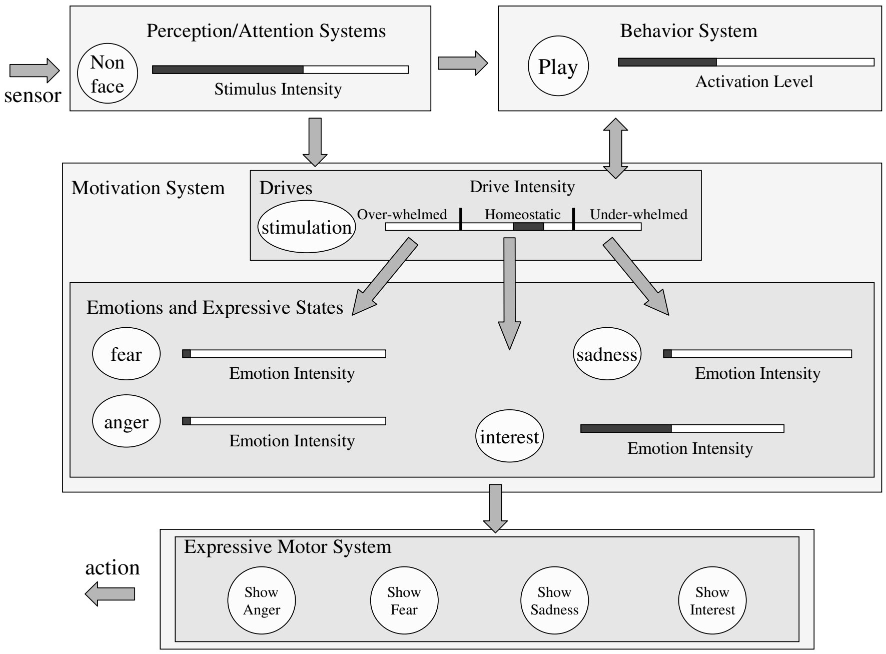

Figure 10: Portions of the behavior engine active during the non-face stimuli experiments. Non-face stimuli activate the `play` behavior, which is potentiated by the `stimulation drive`. The `stimulation drive` acts upon the `emotions` `fear`, `sadness`, `anger`, and `interest`.

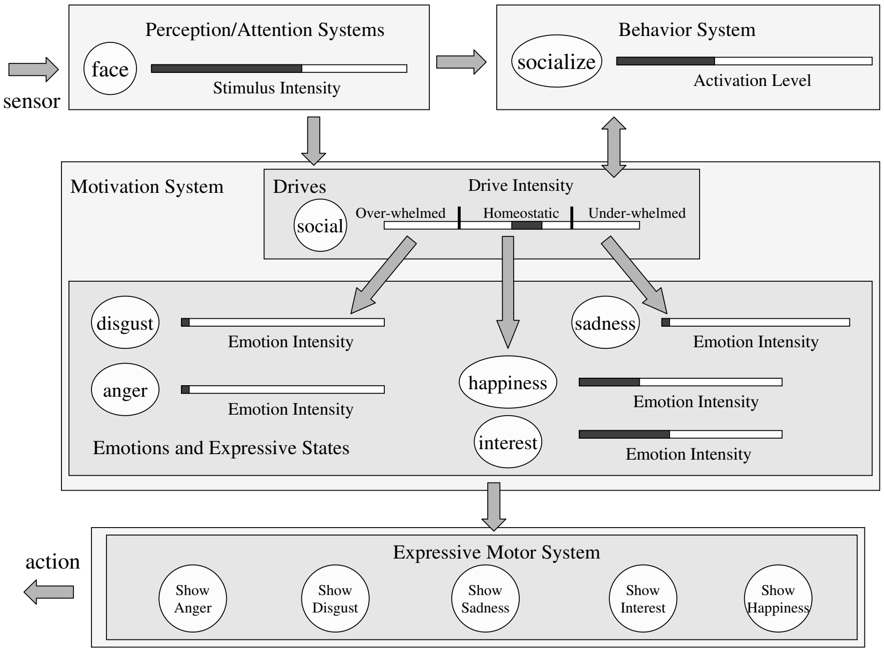

Figure 11: Portions of the behavior engine active during the face stimuli experiments. Face stimuli activate the `socialize` behavior, which is potentiated by the `social drive`. The `social drive` acts upon the `emotions` `disgust`, `anger`, `sadness`, `happiness`, and `interest`.

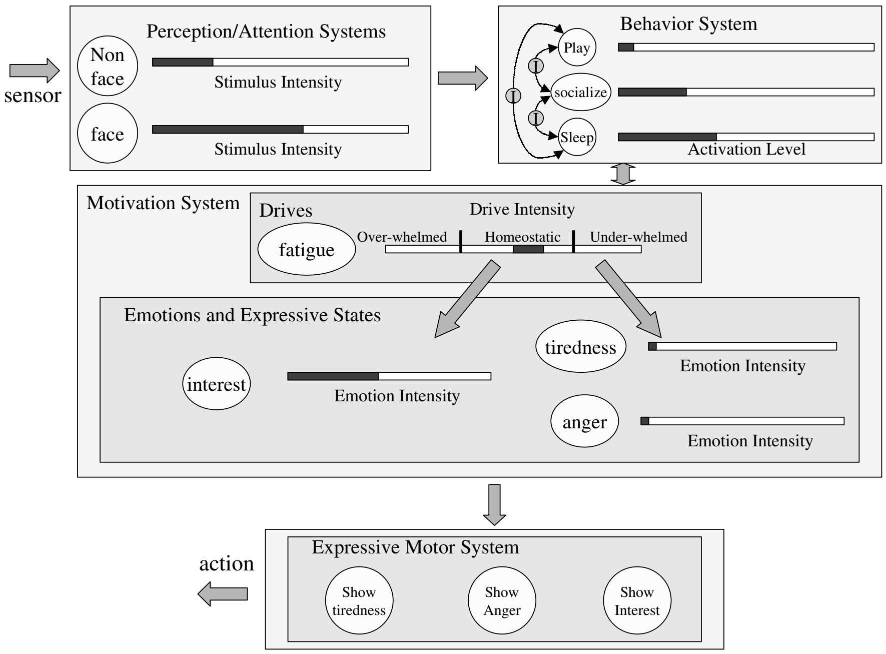

Figure 12: Portions of the behavior engine active in the overstimulation experiments. Both face and non-face stimuli inhibit the `sleep` behavior, which is potentiated by the `fatigue drive`. The `fatigue drive` acts upon the `emotions` `interest`, `tiredness`, and `anger`.

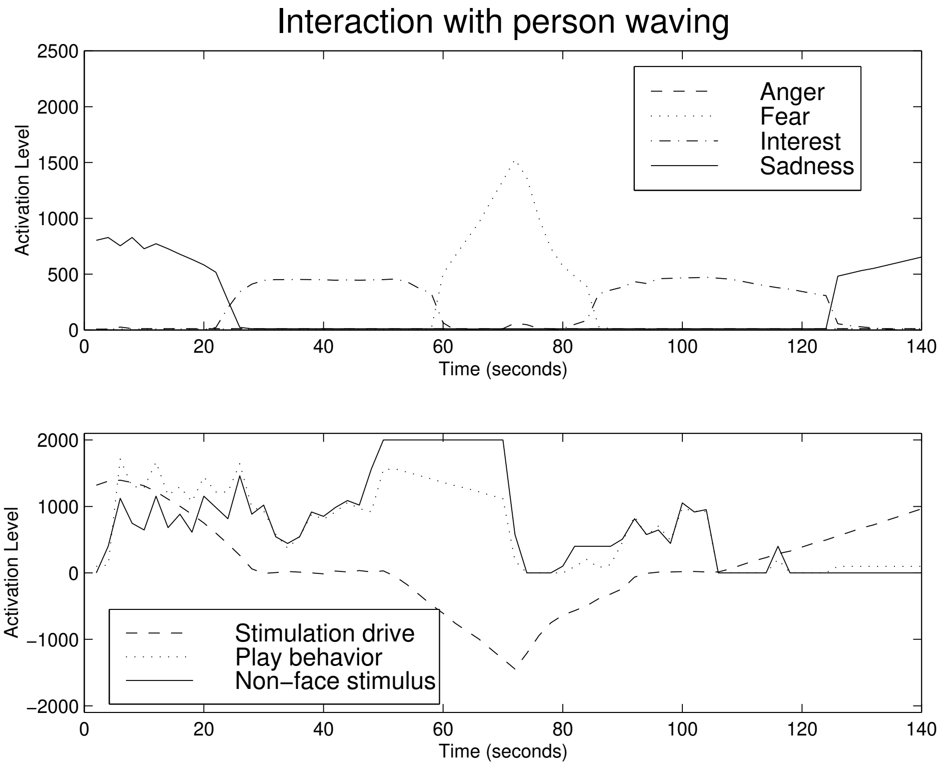

Figure 13: Experimental results for the robot interacting with a person waving. The top chart shows the activation levels of the `emotions` involved in this experiment as a function of time. The bottom chart shows the activation levels of the `drives`, behaviors, and percepts relevant to this experiment. So long as the waving continues at a reasonable intensity, the robot remains `interested`. When the stimulus intensity becomes too great, the robot begins to show `fear`.

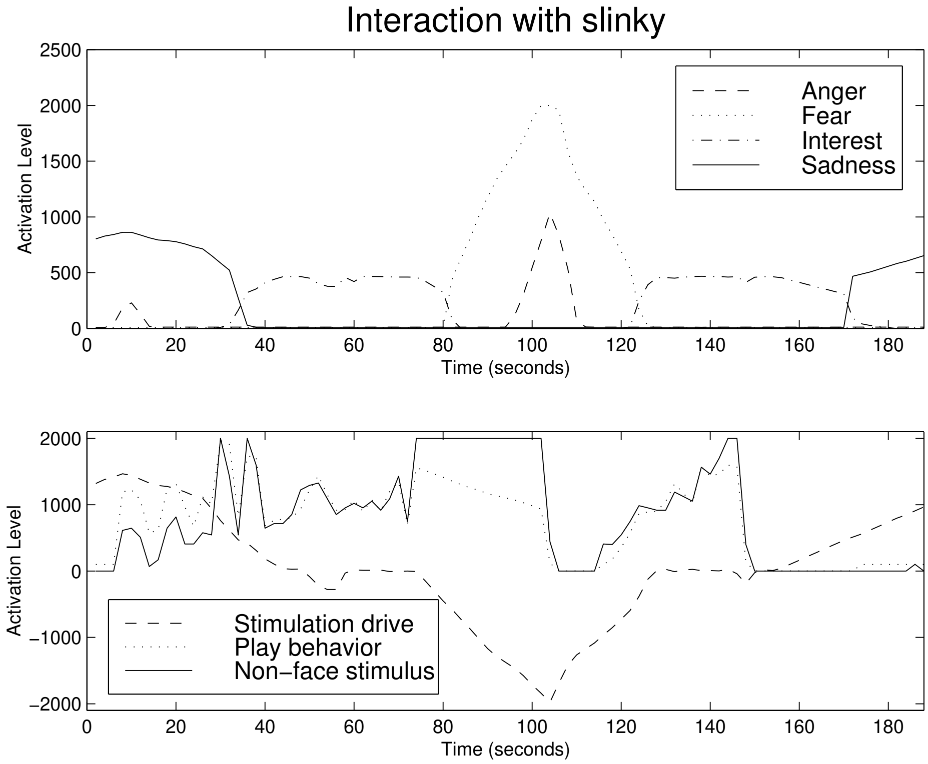

Figure 14: Experimental results for the robot interacting with a toy slinky. So long as the slinky continues to move at a reasonable intensity, the robot remains `interested`. When the stimulus intensity becomes too great, the robot begins to show `fear`, which eventually leads to `anger`.

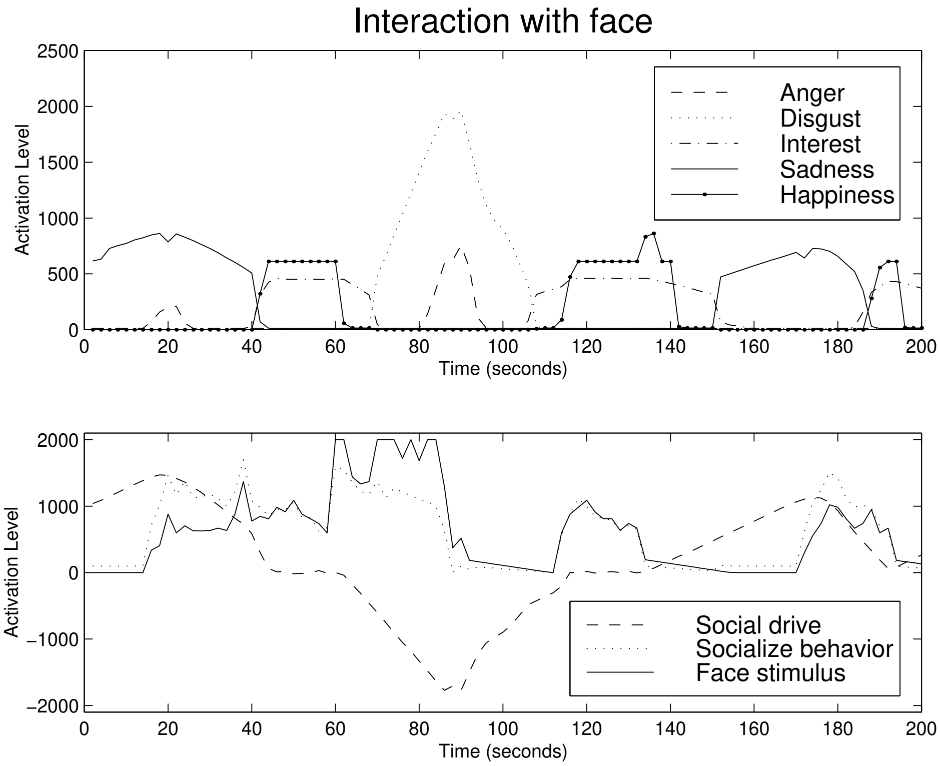

Figure 15: Experimental results for the robot interacting with a person's face. When the face is present, the robot looks `interested` and `happy`. When the face begins to move too violently, the robot begins to show `disgust`, which eventually leads to `anger`. Note that the robot reacts differently to a social stimulus (in this case, a face) than to the previous non-social stimuli.

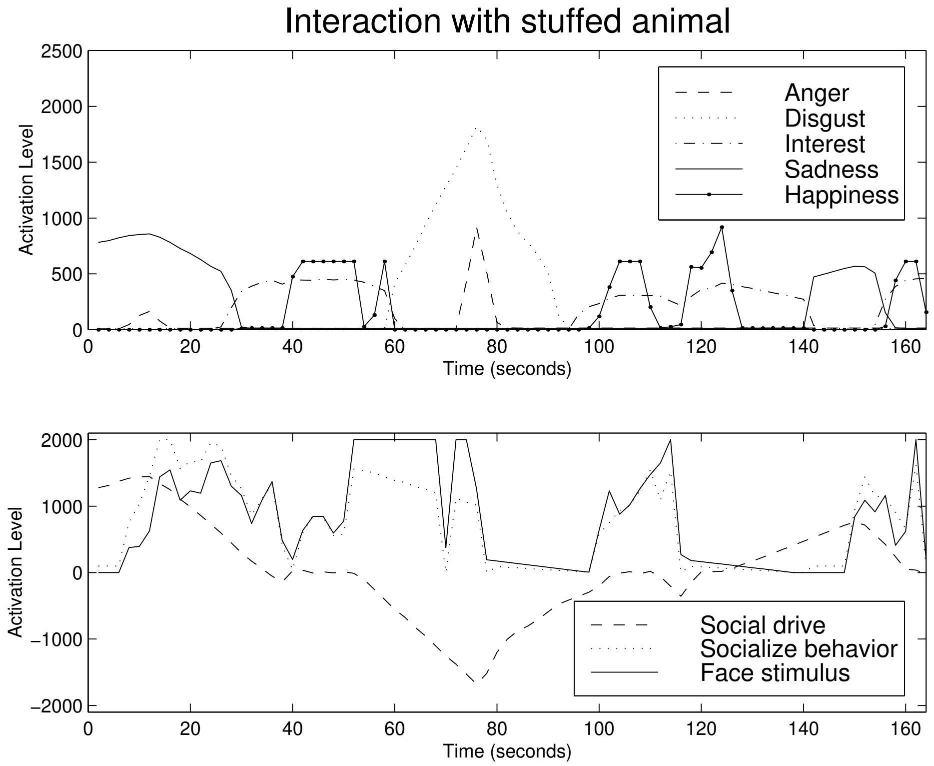

Figure 16: Experimental results for the robot interacting with a toy stuffed animal. The perceptual system recognizes the face of the toy, and the stimulus is classified as a social object. When the face is present, the robot looks `interested` and `happy`. When the face begins to move too violently, the robot begins to show `disgust`, which eventually leads to `anger`.

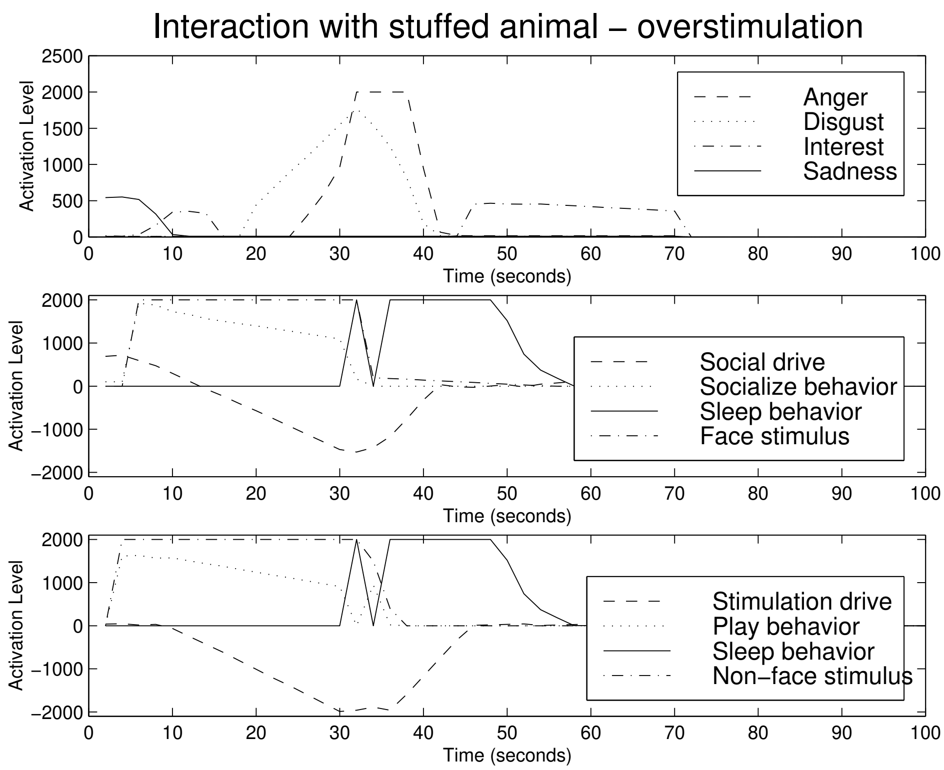

Figure 17: Further experimental results for the robot interacting with a toy stuffed animal. In this case, the experimenter continues to stimulate the robot by moving the stuffed animal even after the robot displays both `disgust` and `anger`. The `sleep` behavior is then activated as an extreme measure to block out stimulation. The `sleep` behavior restores the `drives` and `emotions` to homeostatic levels before allowing the robot to "wake-up."

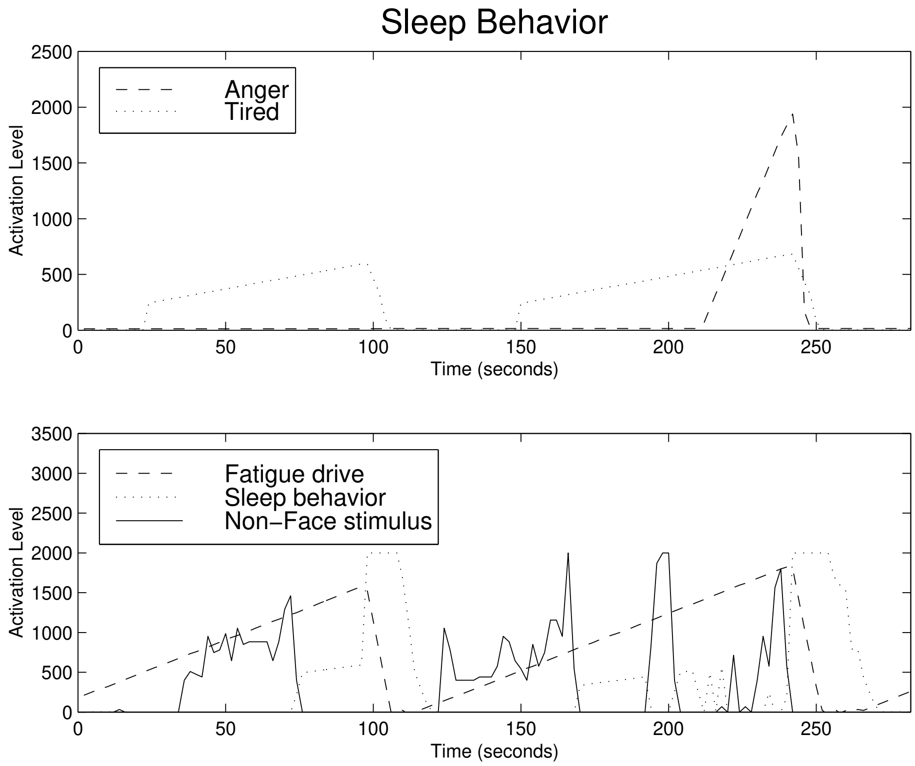

Figure 18: Experimental results for long-term interactions of the `fatigue` drive and the `sleep` behavior. The `fatigue` drive continues to increase until it reaches an activation level that potentiates the `sleep` behavior. If there is no other stimulation, this will allow the robot to activate the `sleep` behavior.