A context-dependent attention system for a social robot

Content Areas: robotics, cognitive modeling, perception
Tracking Number: A770

Abstract

This paper presents part of an on-going project to integrate perception, attention, drives, emotions, behavior arbitration, and expressive acts for a robot designed to interact socially with humans. We present the design of a visual attention system based on a model of human visual search behavior from Wolfe (1994). The attention system integrates perceptions (motion detection, color saliency, and face popouts) with habituation effects and influences from the robot's motivational and behavioral state to create a context-dependent attention activation map. This activation map is used to direct eye movements and to satiate the drives of the motivational system.

1 Introduction

Socially intelligent robots provide both a natural human-machine interface and a mechanism for bootstrapping more complex behavior. However, social skills often require complex perceptual, motor, and cognitive abilities [Brooks *et al.*, 1998]. Our research has focused on a developmental approach to building socially intelligent robots that utilize natural human social cues to interact with and learn from human caretakers.

This paper discusses the construction of one necessary sub-system for social intelligence: an attention system. To provide a basis for more complex social behaviors, an attention system must direct limited computational resources and select among potential behaviors by combining perceptions from a variety of modalities with the existing motivational and behavioral state of the robot. We present a robotic implementation of an attention system based upon models of human attention and visual search. We further outline the ways in which this model interacts with existing perceptual, motor, motivational, and behavioral systems.

Our implementation is based upon Wolfe's model of human visual attention and visual search [Wolfe, 1994]. This model integrates evidence from Treisman [1985], Julesz [1988], and others to construct a flexible model of human visual search behavior. In Wolfe's model, visual stimuli are filtered by broadly-tuned "categorical" channels (such as color and orientation) to produce *feature maps* with activation based upon both local regions (bottom-up) and task demands (top-down). The feature maps are combined by a weighted sum to produce an *activation map*. Limited cognitive and motor resources are distributed in order of decreasing activation. This model has been tested in simulation, and yields results that are similar to those observed in human subjects [Wolfe, 1994]. In this paper we do not attempt to match human performance (a task that is difficult with current component technology), but rather require only that the robotic system perform enough like a human that it is capable of maintaining a normal social interaction. Our implementation is similar to other models based in part on Wolfe's work [Itti *et al.*, 1998; Hashimoto, 1998; Driscoll *et al.*, 1998], but additionally operates in conjunction with motivational and behavioral models, with moving cameras, and it differs in dealing with habituation issues.

2 Robot Hardware

Our robotic platform consists of a stereo active vision system (described in [Scassellati, 1998a]) augmented with facial features for emotive expression. The robot, called Kismet and shown in Figure 1, is able to show expressions (analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise) which are easily interpreted by an untrained human observer. The platform has four degrees of freedom in the vision system; each eye has an independent vertical axis of rotation (pan) and the eyes share a joint horizontal axis of rotation (tilt). Kismet also has fifteen degrees of freedom in facial features, including eyebrows, ears, eyelids, lips, and a mouth. Each eyeball has an embedded color CCD camera with a 5.6 *mm* focal length.

The active vision platform is attached to a parallel network of eight 50M Hz digital signal processors (Texas Instruments TMS320C40). The DSP network serves as the sensory processing engine and implements the bulk of the robot's perception and attention systems. A pair of Motorola 68332-based microcontrollers are also connected to the robot. One controller implements the motor sys-

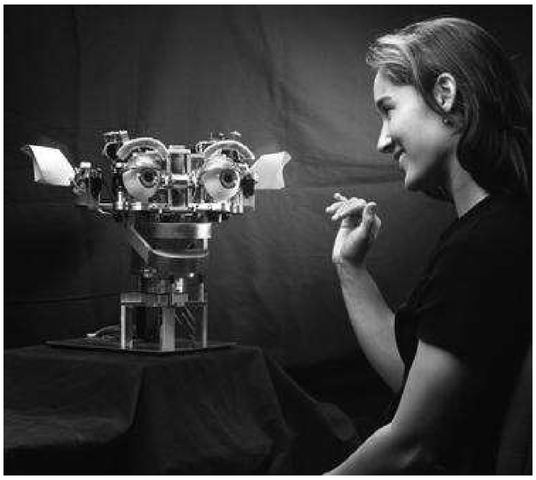

Figure 1: Kismet, a robot designed to interact socially with humans. Kismet has an active vision system and can display a variety of facial expressions.

tem for driving the robot's facial motors. The other controller implements the motivational system (emotions and drives) and the behavior system. The microcontrollers communicate with the DSP network through a dual-ported RAM.

3 Perceptual Systems

Our current perceptual systems focus on the pre-attentive, massively parallel stage of human vision that processes information about basic visual features (color, motion, various depth cues, etc.). The implementation described here focuses on three such pre-attentive processes: color, motion, and face pop-outs. In terms of the model from Wolfe [1994], our implementation contains the bottom-up feature maps, which represent the inherent saliency of a specific image property for each point in the visual scene, and incorporates top-down influences from motivational and behavioral sources.

The video signal from each of Kismet's cameras is digitized by one of the DSP nodes with specialized frame grabbing hardware. The image is then subsampled and averaged to an appropriate size. For these initial tests, we have used an image size of 64×64, which allows us to complete all of the processing in near real-time. To minimize latency, each feature map is computed by a separate DSP processor (each of which also has additional computational task load). All of the feature detectors discussed here can operate at multiple scales.

3.1 Color Saliency Feature Maps

One of the most basic and widely recognized visual feature is color. Our models of color saliency are drawn from the complimentary work on visual search and attention from Itti, Koch, and Niebur [1998]. The incoming video stream contains three 8-bit color channels (r, g, and b) which are transformed into four color-opponency channels (r', g', b', and y'). Each input color channel is first normalized by the luminance l (a weighted average of the three input color channels):

$$r_n = \frac{255}{3} \cdot \frac{r}{l} \qquad g_n = \frac{255}{3} \cdot \frac{g}{l} \qquad b_n = \frac{255}{3} \cdot \frac{b}{l} \tag{1}$$

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

$$r' = r_n - (g_n + b_n)/2 \tag{2}$$

$$g' = g_n - (r_n + b_n)/2 \tag{3}$$

$$b' = b_n - (r_n + g_n)/2 \tag{4}$$

$$y' = \frac{r_n + g_n}{2} - b_n - \|r_n - g_n\| \tag{5}$$

The four opponent-color channels are clamped to 8-bit values by thresholding. While some research seems to indicate that each color channel should be considered individually [Nothdurft, 1993], we choose to maintain all of the color information in a single feature map to simplify the processing requirements (as does Wolfe [1994] for more theoretical reasons). The maximum of the four opponent-color values is computed and then smoothed with a uniform 5×5 field to produce the output color saliency feature map. This smoothing serves both to eliminate pixel-level noise and to provide a neighborhood of influence to the output map, as proposed by Wolfe [1994]. A single DSP node computes these computations and forwards the resulting feature map both to the attention process at a rate of 20-25 Hz. The processor produces a pseudo-color image by scaling the luminance of the original image by the output saliency while retaining the same relative chrominance (as shown in Figure 2).

3.2 Motion Saliency Feature Maps

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

$$M_{raw} = \mathcal{T}(\|I_t - I_{t-1}\|) \tag{6}$$

This raw motion map is then smoothed with a uniform 7×8 field. While using a 5×5 field would have maintained consistency with both Wolfe's model and the color saliency feature map, using a slightly larger field size allows us to use the output of the motion saliency map as a pre-filter to the face detection routine, which has optimized performance in prior tests by a factor of 3 [Scassellati, 1998b]. The motion saliency feature map is computed at 25-30 Hz by a single DSP processor node and forwarded both to the attention process and the VGA display.

3.3 Face Pop-Out Feature Maps

While form and size are part of Wolfe's original model, we have extended the concept to include other known

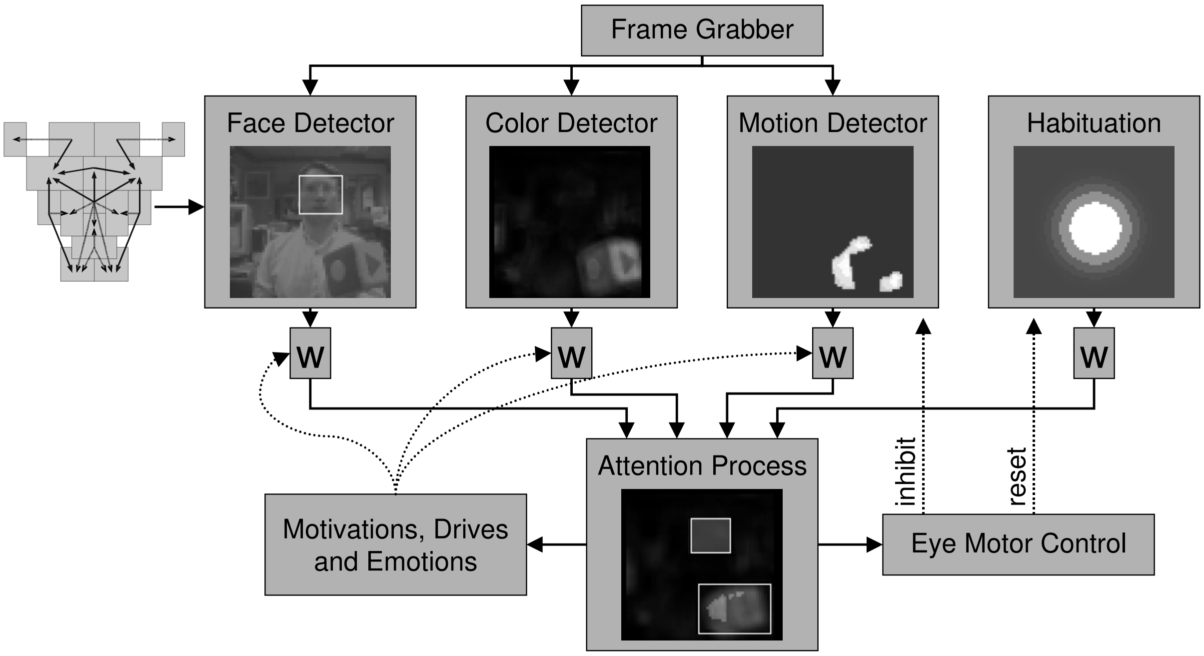

Figure 2: Overview of the attention system. A variety of visual feature detectors (color, motion, and face detectors) combine with a habituation function to produce an attention activation map. The attention process influences eye control and the robot's internal motivational and behavioral state, which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

pop-out features that have social relevance such as faces. Our face detection techniques are designed to identify locations that are likely to contain a face, not to verify with certainty that a face is present in the image. The face detector is based on the ratio-template technique developed by Sinha [1996], and has been previously reported [Scassellati, 1998b]. 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]. Ratio templates also offer multiple levels of biological plausibility; templates can be either hand-coded or learned adaptively from qualitative image invariants [Sinha, 1996].

A ratio template is composed of regions and relations, as shown to the left of the face detector in Figure 2. For each target location in the grayscale peripheral image, a template comparison is performed using a special set of comparison rules. The set of regions is convolved with a 14×16 image patch around a pixel location 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). The number of satisfied relations serves as the match score for a particular location; the more relations that are satisfied the more likely that a face is located there. In Figure 2, each arrow indicates a relation, with the head of the arrow denoting the second region (the denominator of the ratio).

The ratio template algorithm has been shown to be reasonably invariant to changes in illumination and slight rotational changes [Scassellati, 1998b]. The ratio template algorithm processes video streams in real time using optimization and pre-filtering techniques, and the system has been tested on a variety of lighting conditions and subjects. The algorithm can operate on each level of an image pyramid in order to detect faces at multiple scales. In the current implementation, due to limited processing capability, we elected to process only a single scale for faces. Applied to a 64×64 image from Kismet's cameras, the 14×16 ratio template finds faces in a range of approximately 3-6 feet from the robot. This range was suitable for our current investigations of face-to-face social interactions, and could easily be expanded with additional processors. The implemented face detector operates at approximately 15-20 Hz.

4 Behaviors and Motivations

In previous work, Breazeal and Scassellati [1998] presented how the design of Kismet's motivation and behavior systems enable it to socially interact with a human while regulating the intensity of the interaction via expressive displays. For the purposes of this paper, we present only those aspects of these systems which bias the robot's attention (see Figure 3).

Perceptual stimuli are classified into *social* stimuli (i.e. people, which move and have faces) which satisfy a drive

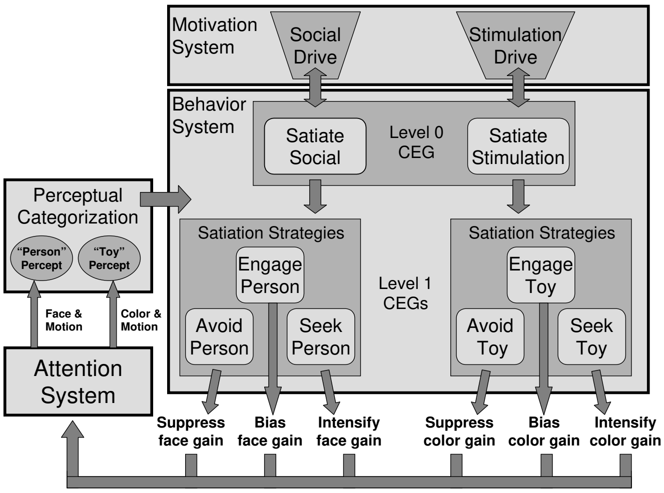

Figure 3: Schematic of motivations and behaviors relevant to attention. See text for details.

to be social and *non-social* stimuli (i.e. toys, which move and are colorful) which satisfy a drive to be stimulated by other things in the environment.

For each drive, there is a desired operation point, and an acceptable bounds of operation around that point (the *homeostatic regime*). As long as a drive is within the homeostatic regime, that corresponding need is being adequately met. Unattended, drives drift toward an under-stimulated regime. Excessive stimulation (too many stimuli or stimuli moving too quickly) push a drive toward an over-stimulated regime.

The robot's drives influence behavior selection by preferentially passing activation to select behaviors. By doing so, the robot is more likely to activate behaviors that serve to restore its drives to their homeostatic regimes. The top level (level 0) of the behavior system consists of a single cross-exclusion group (CEG) containing two behaviors: `satiate social` and `satiate stimulation`. Each behavior is viewed as a self-interested, goal-directed process. Within a CEG, behaviors compete for activation in a winner-take-all scheme based upon perceptual factors, motivational factors, and its own behavioral persistence. Competition between behaviors at the top level represents selection at the *task* level. By organizing the top level behaviors in this fashion, the robot can only act to restore one drive at a time. This is reasonable since the satiating stimuli for each drive are mutually exclusive and require different behaviors. Specifically, whenever the `satiate social` behavior wins, the robot's task is to do what it must to restore the `social` drive, and when the `satiate stimulation` behavior wins, the robot's task is to do what it must to restore the `stimulation` drive.

Each behavior node of the top level CEG has a child CEG (level 1) associated with it. Once a level 0 behavior wins the competition, it activates its child CEG at level 1. Subsequently, the behaviors within the active level 1 CEG compete for activation. Competition between behaviors within the active level 1 CEG represents competition at the *strategy* level. Each behavior has its own distinct conditions for becoming relevant and winning the competition. For instance, the `avoid person` behavior is the most relevant when the robot's social drive is in the overwhelmed regime and a person is stimulating the robot too vigorously. The goal of this behavior is to reduce the intensity of stimulation. If successful, the `social` drive will be restored to the homeostatic regime. Similarly, the goal of the `seek person` behavior is to acquire a social stimulus of reasonable intensity. If successful, this will serve to restore the `social` drive from the under-stimulated regime. The `engage person` behavior is active by default (i.e. the `social` drive is already in the homeostatic regime and the robot is receiving a good quality stimulus).

5 Attention System

The attention system must combine the various effects of the perceptual input with the existing motivational and behavioral state of the robot both to direct limited computational resources and to select among potential behaviors. Figure 2 shows an overview of the attention system.

5.1 Combining Perceptual Inputs

Each of the feature maps contains an 8-bit value for each pixel location which represents the relative presence of that visual scene feature at that pixel. The attention process combines each of these feature maps using a weighted sum to produce an attention activation map (using the terminology of Wolfe [1994]). The gains for each feature map default to values of 200 for color, 40 for motion, and 50 for face detection. The attention activation map is thresholded to remove noise values, and normalized by the sum of the gains. Connected object regions are extracted using a grow-and-merge procedure with 8-connectivity. To further combine related regions, any regions whose bounding boxes have a significant overlap are also merged.

Statistics on each region are collected, including the centroid, bounding box, area, average attention activation score, and average score for each of the feature maps in that region. The tagged regions that have an area in excess of 30 pixels are sorted based upon their average attention activation score. The attention process provides the top three regions to both the eye motor control system and the behavior and motivational systems. The eye motor control system uses the centroid of the most salient regions to determine where to look next. The top-down processes use the attention activation score and the individual feature map scores of the most salient region to determine which of the drives and behaviors will become activated.

5.2 Attention Drives Eye Movement

The eye motor control process acts on the data from the attention process to center the eyes on an object within the visual field. Our current implementation uses a static linear mapping between image position and eye position, which has been sufficient for our initial investigations. We are currently in the process of converting to a self-calibrated system that learns the sensori-motor mapping for foveation similar to that described by Scassellati [1998a].

Each time that the eyes move, the eye motor process sends two signals. The first signal inhibits the motion detection system for approximately 600 msec, which prevents self-motion from appearing in the motion feature map. The second signal resets the habituation state, which is described below.

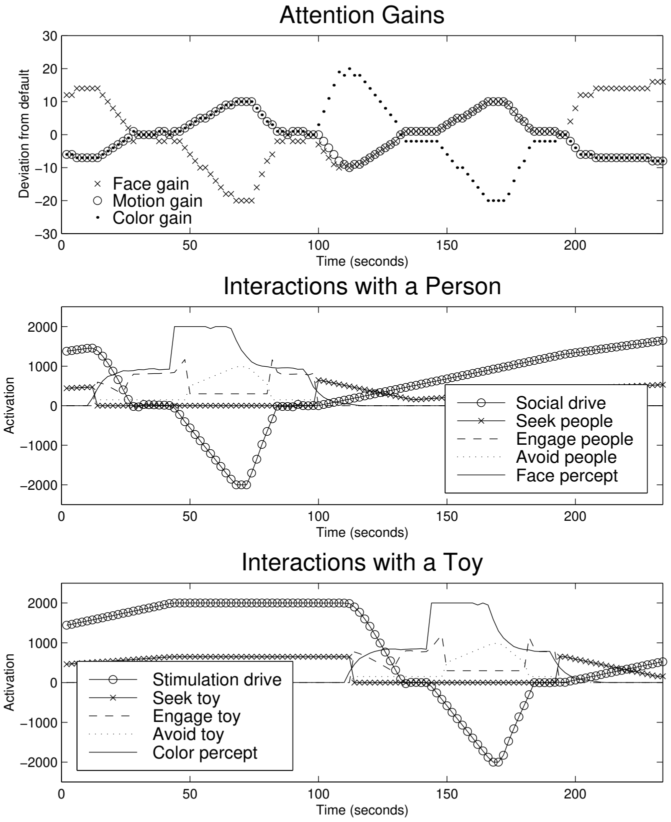

Figure 4: Changes of the face, motion, and color gains from top-down motivational and behavioral influences (top). When the `social` drive is activated by face stimuli (middle), the face gain is influenced by the `seek people` and `avoid people` behaviors. When the `stimulation` drive is activated by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

5.3 Habituation

For our robot, the current object under consideration is always the object that is in the center of the visual field.[1] The habituation function can be viewed as a feature map that initially maintains eye fixation by increasing the saliency of the center of the field of view and slowly decays the saliency values of central objects until a salient off-center object causes the eyes to move. The habituation function is a Gaussian field $G(x, y)$ centered in the field of view with peak amplitude of 255 (to remain consistent with the other 8-bit values) and $\theta = 50$ pixels. It is combined linearly with the other feature maps using the weight

$$w = W \cdot max(-1, 1 - \Delta t/\tau) \qquad (7)$$

where w is the weight, Δt is the time since the last habituation reset, τ is a time constant, and W is the maximum habituation gain. Whenever the eyes move, the habituation function is reset, forcing w to W and amplifying the saliency of central objects until a time τ when $w = 0$ and there is no influence from the habituation map. As time progresses, w decays to a minimum value of $-W$ which suppresses the saliency of central objects. In the current implementation, we use a value of $W = 10$ and a time constant $\tau = 5$ seconds.

The entire attention process (with habituation) operates at 10-25 Hz on a single DSP processor node. The speed varies with the number of attention activation pixels that pass threshold for region growing. While this code could be optimized further, rates above 10 Hz are not necessary for our current purposes.

5.4 Motivations and Behaviors Influence Feature Map Gains

Kismet's drives and behaviors bias the attentional gains based on the current internal context to preferentially attend to behaviorally relevant stimuli. Behaviors that satiate the `stimulation` drive influence the color saliency gain because color is characteristic of toys. Similarly, the face saliency gain is adjusted when the robot is tending to its `social` drive. Active level 1 behaviors influence attentional gains in proportion to the intensity of the associated drive.

As shown in Figure 3, the face gain is enhanced when the `seek people` behavior is active and is suppressed when the `avoid people` behavior is active. Similarly, the color gain is enhanced when the `seek toys` behavior is active, and suppressed when the `avoid toys` behavior is active. Whenever the `engage people` or `engage toys` behaviors are active, the face and color gains are restored to their default values, respectively. Weight adjustments are constrained such that the total sum of the weights remains constant at all times. Figure 4 illustrates how the face, motion, and color gains are adjusted as a function of drive intensity, the active level 1 behavior, and the nature and quality of the perceptual stimulus.

6 Results and Evaluation

Top-down gain adjustments combine with bottom-up habituation effects to bias the robot's gaze preference (see

[1]This is extremely relevant on our other robotic platforms which have a second camera that captures a high resolution foveal image.

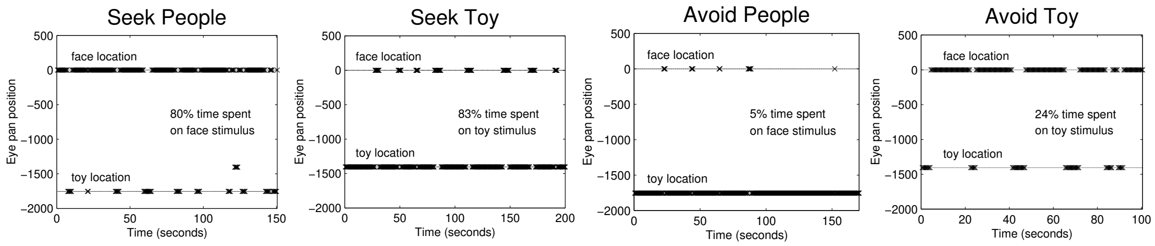

Figure 5: Preferential looking based on habituation and top-down influences. When presented with two salient stimuli (a face and a brightly colored toy), the robot prefers to look at the stimulus that has behavioral relevance. Habituation causes the robot to also spend time looking at the non-preferred stimulus.

Figure 5). When the `seek people` behavior is active, the face gain is enhanced and the robot prefers to look at a face over a colorful toy. The robot eventually habituates to the face stimulus and switches gaze briefly to the toy stimulus. Once the robot has moved its gaze away from the face stimulus, the habituation is reset and the robot rapidly re-acquires the face. In one set of behavioral trials when `seek people` was active, the robot spent 80% of the time looking at the face. A similar affect can be seen when the `seek toy` behavior is active — the robot prefers to look at a toy over a face 83% of the time.

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

7 Future Work

In this paper we have demonstrated an attentional system that combines bottom-up perceptions and habituation effects with top-down behavioral and motivational influences. This results in a system that directs eye gaze based on current task demands. In the future, we intend to construct a richer set of perceptual inputs (depth, orientation, and texture) and motor responses (smooth pursuit tracking, vergence, and vestibulo-ocular reflex). We are also currently combining this system with expressive behaviors to facilitate social interaction with a human.

References

[Breazeal and Scassellati, 1998] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, 1998. To appear.

[Brooks *et al.*, 1998] Rodney A. Brooks, Cynthia Breazeal (Ferrell), Robert Irie, Charles C. Kemp, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. Alternative essences of intelligence. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98)*, 1998.

[Driscoll *et al.*, 1998] Joseph A. Driscoll, Richard Alan Peters II, and Kyle R. Cave. A visual attention network for a humanoid robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-98)*, 1998.

[Hashimoto, 1998] S. Hashimoto. Humanoid robots in Waseda University - Hadaly-2 and WABIAN. In *IARP First International Workshop on Humanoid and Human Friendly Robotics*, Tsukuba, Japan, 1998.

[Itti *et al.*, 1998] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[Julesz and Krose, 1988] B. Julesz and B. Krose. Features and spatial filters. *Nature*, 333:302–303, 1988.

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

[Scassellati, 1998a] Brian Scassellati. A binocular, foveated active vision system. Technical Report 1628, MIT Artificial Intelligence Lab Memo, 1998.

[Scassellati, 1998b] Brian Scassellati. Finding eyes and faces with a foveated vision system. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98)*, 1998.

[Sinha, 1996] Pawan Sinha. *Perceiving and recognizing three-dimensional forms*. PhD thesis, Massachusetts Institute of Technology, 1996.

[Treisman, 1985] A. Treisman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31:156–177, 1985.

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