Social Constraints on Animate Vision

Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, Brian Scassellati, Paulina Varchavskaia,
MIT Artificial Intelligence Laboratory

THE CHALLENGE OF INTERACTING WITH HUMANS CONSTRAINS ROBOTS' PHYSICAL APPEARANCE, ENVIRONMENTAL PERCEPTION, AND BEHAVIORAL ORGANIZATIONHOW ROBOTS APPEAR, MOVE, PERCEIVE, AND BEHAVE. THE AUTHORS PRESENT A VISUAL-MOTOR SYSTEM THAT NEGOTIATES BETWEEN THESE CONSTRAINTS TO FACILITATE ROBOT FUNCTIONALITY IN A SOCIAL ENVIRONMENT.

In a social context, people constantly read each other's actions as clues to attention and motivation. If an anthropomorphic robot can evoke similar treatment, interaction with a human can be very smooth and intuitive. But this will only hold true if the robot can structure its actions so that meaning assigned to them by a naïve observer is in fact veridical. This places far-reaching constraints on the robot's design, particularly on the organization of its visual-motor system.

Human eye movements have a high communicative value. A person's gaze direction and eye movement (staring *versus* glancing, making and breaking eye contact) can convey important information during social interactions. Modeling an anthropomorphic robot's eye movements after humans makes its behavior easily understood as analogous to human behavior in similar circumstances (see Figure 1).

There are other advantages to following a biological model. Researchers can use existing data and proposed models for the human visual system organization. Another advantage is integrating action, perception, attention and other cognitive capabilities, which makes the system more flexible and reliable. Adding additional perceptual capabilities and additional constraints between behavioral and perceptual modules can increase the behavioral relevance while limiting the computational requirements.[2] For example, combining the tracking mechanism with a visual attention system helps identifying objects that are behaviorally relevant and worth tracking.

Physical form

Currently, our most sophisticated robot in terms of visual-motor behavior is Kismet. This robot is an active vision head augmented with expressive facial features (see Figure 2). We designed Kismet to receive and send human-like social cues to a caregiver, who can regulate its environment and shape its experiences as a parent would for a child.[3] Kismet has three degrees of freedom to control gaze direction, three degrees of freedom for its neck, and fifteen degrees of freedom in other expressive facial components (such as ears and eyelids). To perceive its caregiver Kismet uses a microphone, worn by the caregiver, and four color CCD cameras. The positions of the neck and eyes are important both for expressive postures and for directing the cameras towards behaviorally relevant stimuli.

The cameras in Kismet's eyes have high acuity but a narrow field of view. Between the eyes, we fixed two unobtrusive central cameras, each with a wider field of view but correspondingly lower acuity. The reason for using different cameras is that typical visual tasks require both high acuity (for

1094-7167/00/$10.00 © 2000 IEEE

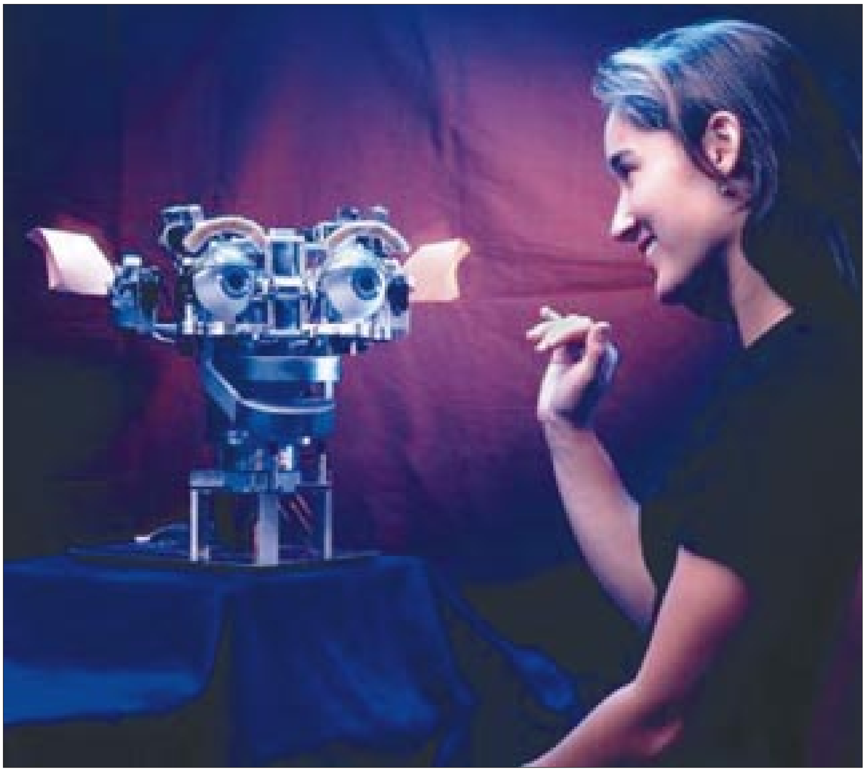

Figure 1. Kismet, a robot capable of conveying intentionality through facial expressions and behavior.[1] Here, the robot's physical state expresses attention to and interest in the human beside it. Another person—for example, the photographer—would expect to have to attract the robot's attention before being able to influence its behavior.

recognition tasks and for controlling precise visually guided motor movements) and a wide field of view (for search tasks, for tracking multiple objects, compensating for involuntary ego-motion, and so on). Biological systems commonly sample part of the visual field at a high enough resolution to support the first task set, and sample the rest of the field at an adequate level to support the second set. We see this in animals with foveate vision, such as humans, where the photoreceptor density is highest at the center and falls off dramatically toward the periphery. Vision researchers mimic this by using specially designed imaging hardware, space-variant image sampling,[4] or by using multiple cameras with different fields of view.

Another of our robots, Cog, follows the human sensing arrangement more closely than does Kismet (see "Humanoid Robots: A New Kind of Tool," by Bryan Adams and his colleagues in this issue).

The designs of our robots are constantly evolving. We add new and reorganize old degrees of freedom, replace or rearrange sensors, and introduce new sensory modalities. The descriptions herein are snapshots of the robots' current state.

System architecture

We have designed our hardware and software control architectures to meet the challenge of real-time visual-signal processing (approaching 30 Hz) with minimal latencies. We've implemented Kismet's vision system on a network of nine 400-MHz commercial PCs running the QNX real-time operating system (see Figure 3). Kismet's motivational and behavioral systems run on four Motorola 68332 processors. We also networked machines running Windows NT and Linux for speech generation and recognition, respectively. Even more so than Kismet's physical form, the control network is rapidly evolving as new behaviors and sensory modalities come online.

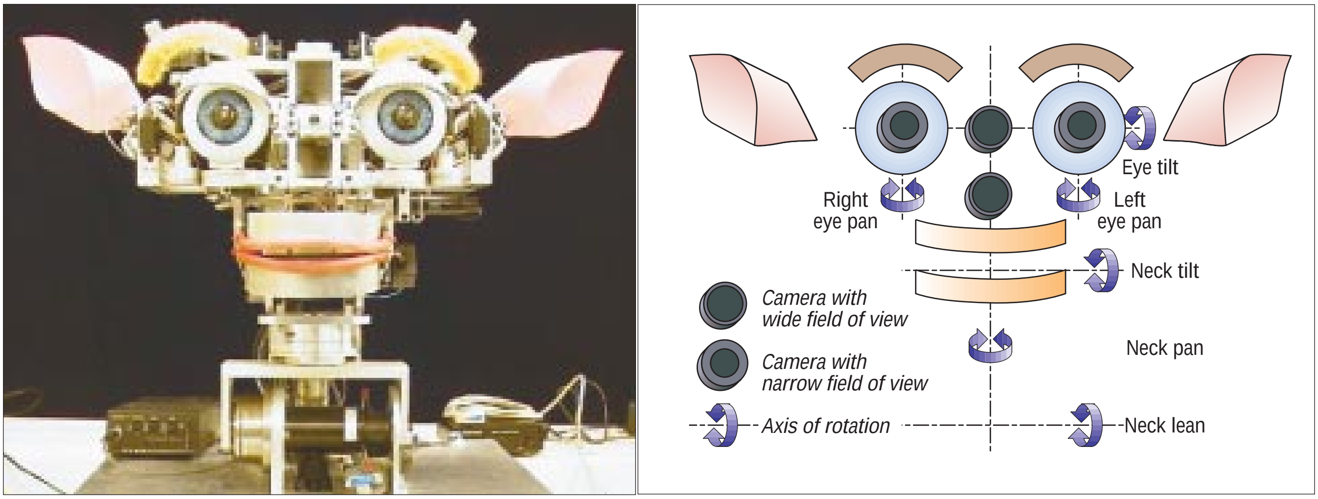

Figure 2: Kismet's expressive features include eyelids, eyebrows, ears, jaw, lips, neck, and eye orientation. The schematic on the right shows the degrees of freedom relevant to visual perception (omitting the eyelids!). The eyes can pan independently along the horizontal, but tilt together along the vertical. The neck can turn the whole head horizontally and vertically, and can also crane forward. Two cameras with narrow fields of view rotate with the eyes. Two central cameras with wide fields of view rotate with the neck. The eye orientation does not affect these cameras.

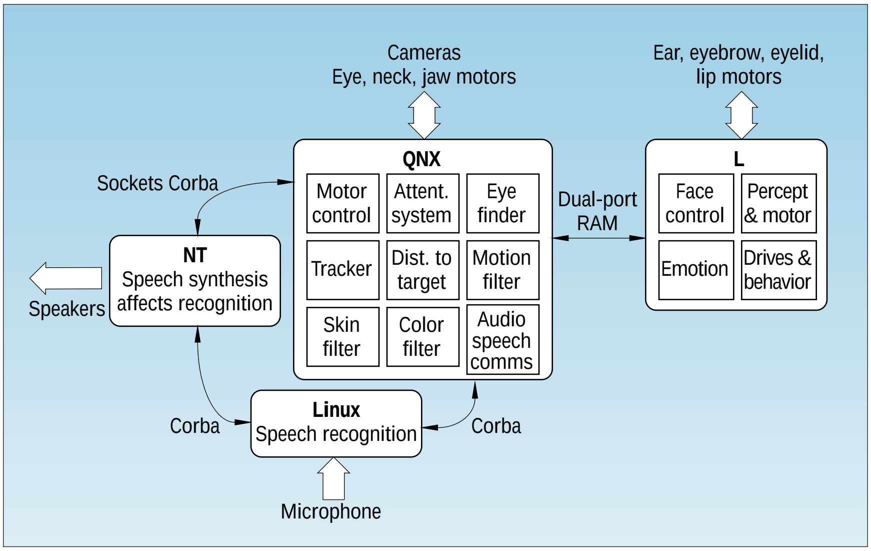

Figure 3. Kismet's system architecture. The motivational and behavioral systems run on four Motorola 68332 microprocessors running L, a multithreaded Lisp developed in our lab. Nine networked PCs running QNX perform the vision processing and eye–neck control.

Pre-attentive visual perception

Human infants and adults naturally find certain perceptual features interesting. Features, such as color, motion, and face-like shapes are very likely to attract our attention.[5] We have implemented various perceptual feature detectors that are particularly relevant to interacting with people and objects. These include low-level feature detectors attuned to quickly moving objects, highly saturated color, and colors representative of skin tones. Figure 4 shows examples of features we have used. The robot also detects looming objects pre-attentively to facilitate a fast reflexive withdrawal.

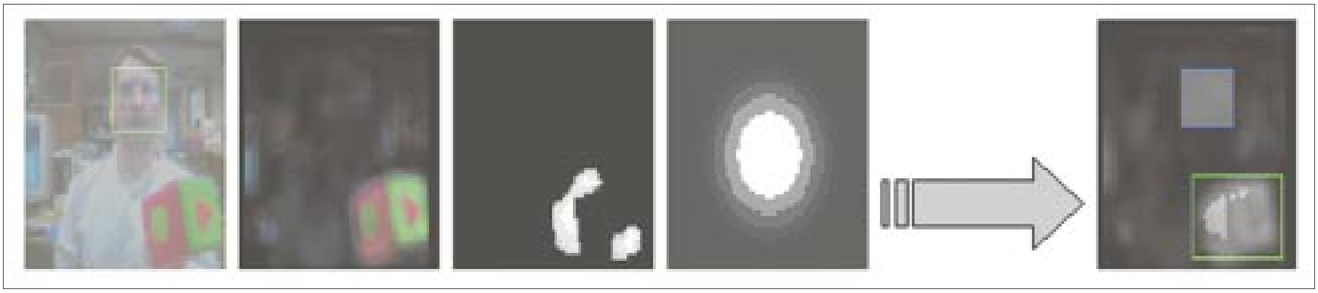

Figure 4. A combination of low-level perceptual stimuli with high-level behavioral and motivational states determines the robot's attention. High-level behavior and motivational influences modulate the low-level features' relative weightings.[6] A sufficiently salient stimulus in any modality can preempt attention, similar to the human response to sudden motion. All else being equal, the robot considers larger objects more salient than smaller ones. The design keeps the robot responsive to unexpected events, while avoiding making it a slave to every whim of its environment. People intuitively provide the right cues to direct the robot's attention (shake object, move closer, wave hand, and so on).

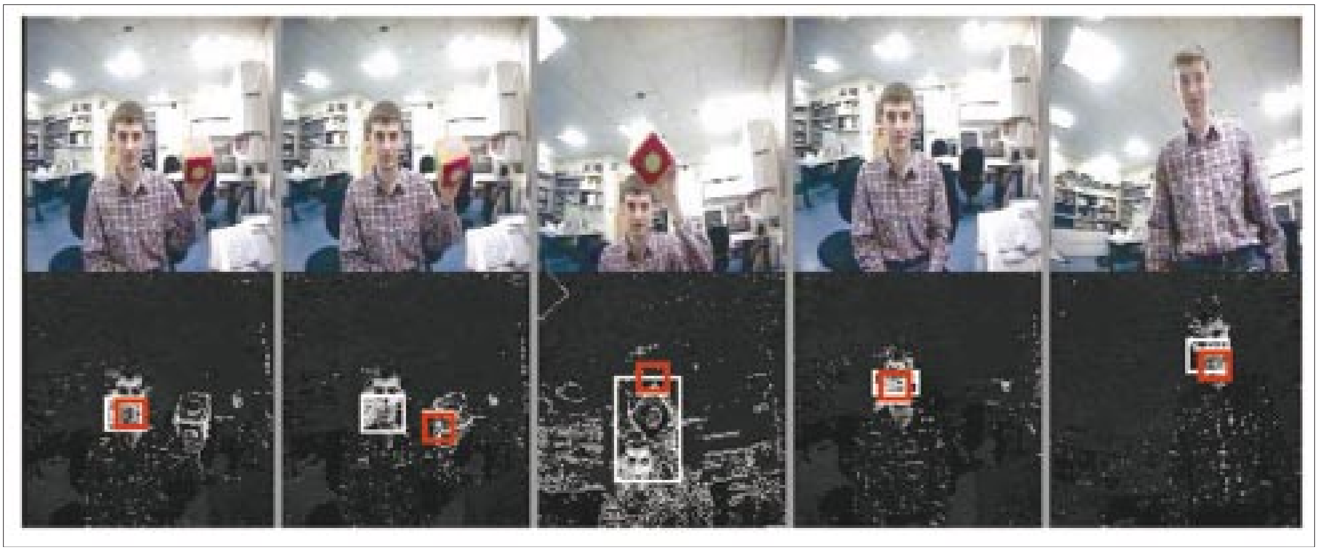

Figure 5. Manipulating the robot's attention. Images on the top row are from Kismet's upper wide camera. Images on the bottom summarize the contemporaneous state of the robot's attention system. Brightness in the lower image corresponds to salience; rectangles correspond to regions of interest. The thickest rectangles correspond to the robot's locus of attention. The robot's motivation here is such that stimuli associated with faces and stimuli associated with toys are equally weighted. In the first pair of images, the robot is attending to a face and engaging in mutual regard. By shaking the colored block, its salience increases enough to cause a switch in the robot's attention. The third pair shows that the head tracks the toy as it moves, giving feedback to the human as to the robot's attention. The eyes are also continually tracking the target more tightly than the neck does. In the fourth pair, the robot's attention switches back to the human's face and tracks it as it moves.

Visual attention

We have implemented Jeremy Wolfe's model of visual search and attention.[7] Our implementation is similar to other models based in part on Wolfe's work,[8] but additionally operates in conjunction with motivational and behavioral models, with moving cameras, and addresses the issue of habituation.

The attention process acts in two parts. The robot produces a single attention map through a weighted combination of various low-level feature detectors (such as color, motion, and shape). This combination allows the robot to select regions that are visually salient and to direct its computational and behavioral resources toward those regions. The attention system also integrates influences from the robot's internal motivational and behavioral systems to bias the selection process. For example, if the robot's current goal is to interact with people, the attention system is biased toward objects that have colors characteristic of skin-tone. The attention system also has mechanisms for habituating to stimuli, thus providing the robot with a primitive attention span. Figure 5 shows an example of the attention system in use, choosing stimuli in a complex scene that are potentially behaviorally relevant. The attention system runs all the time, even when it is not controlling gaze direction, since it determines the perceptual input to which the motivational and behavioral systems respond.

Post-attentive processing

Once the attention system has selected regions within the visual field that are potentially behaviorally relevant, we can apply more intensive computation to these regions. Searching for eyes is one such task. Locating eyes is important to us for engaging in eye contact, and as a reference point for interpreting facial movements and expressions. We currently search for eyes after the robot directs its "foveal" cameras to a candidate region, giving a relatively high-resolution image of the area (Figure 6). Another calculation currently done post-attentively is target distance. The robot estimates this distance using a stereo match between the two central cameras.

Eye movement primitives

We modeled Kismet's visual-motor control after the human ocular-motor system. Humans have foveate vision. The fovea (the center of the retina) has a much higher photoreceptor density than the periphery. This means that to see an object clearly, humans must move their eyes so that the image of the object falls on the fovea. Human eye movement is not smooth. It comprises many quick jumps, called saccades, which rapidly re-orient the eye to project a different part of the visual scene onto the fovea. After a saccade, there is typically a fixation period during which the eyes are relatively stable (but not stationary—they continue to engage in corrective micro-saccades and other small movements). If the eyes fixate on a moving object, they can follow it with an involuntary continuous tracking movement called smooth pursuit (only possible when a moving object is present). Fixation periods typically end after some hundreds of milliseconds, after which a new saccade will occur.[9]

The eyes normally move in lockstep, making equal, conjunctive movements. For a close object, the eyes need to turn toward each other somewhat to correctly image the object on the two eyes' foveae. These disjunctive movements are called vergence, and rely on depth perception (see Figure 7).

Because the eyes are located on the head, they need to compensate for any head movements that occur during fixation. The vestibulo-ocular reflex uses inertial feedback from the vestibular system to keep the eyes' orientation stable as they move. This is a very fast response, but is prone to error accumulation over time. The opto-kinetic response is a slower compensation mechanism that uses a measure of the image's visual slip across the retina to

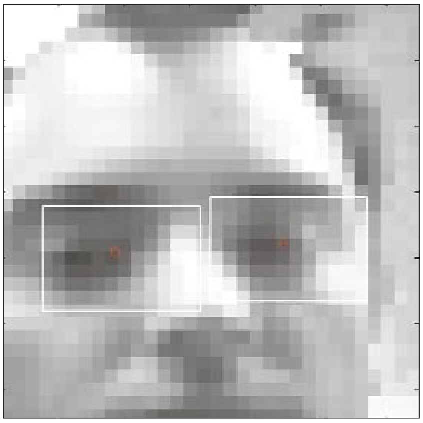

Figure 6. The robot searches for the eyes within a restricted part of its field of view.

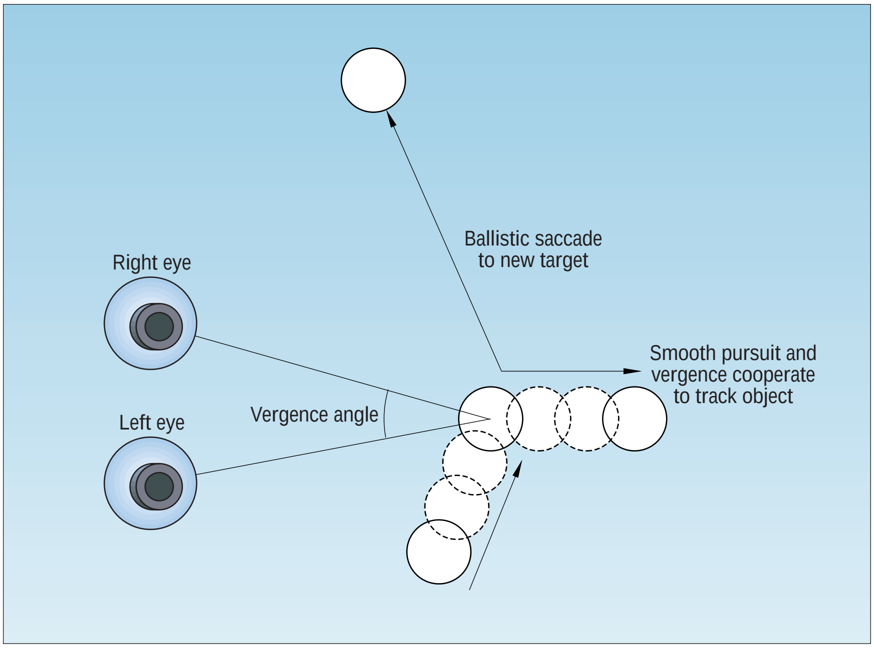

Figure 7. Humans exhibit four characteristic types of eye motion. Saccadic movements are high-speed ballistic motions that center a target in the field of view. Smooth-pursuit movements track a moving object at low velocities. The vestibulo-ocular and opto-kinetic reflexes act to maintain the gaze angle as the head and body move through the world. Vergence movements serve to center an object in both eyes' field of view as the object moves in depth.

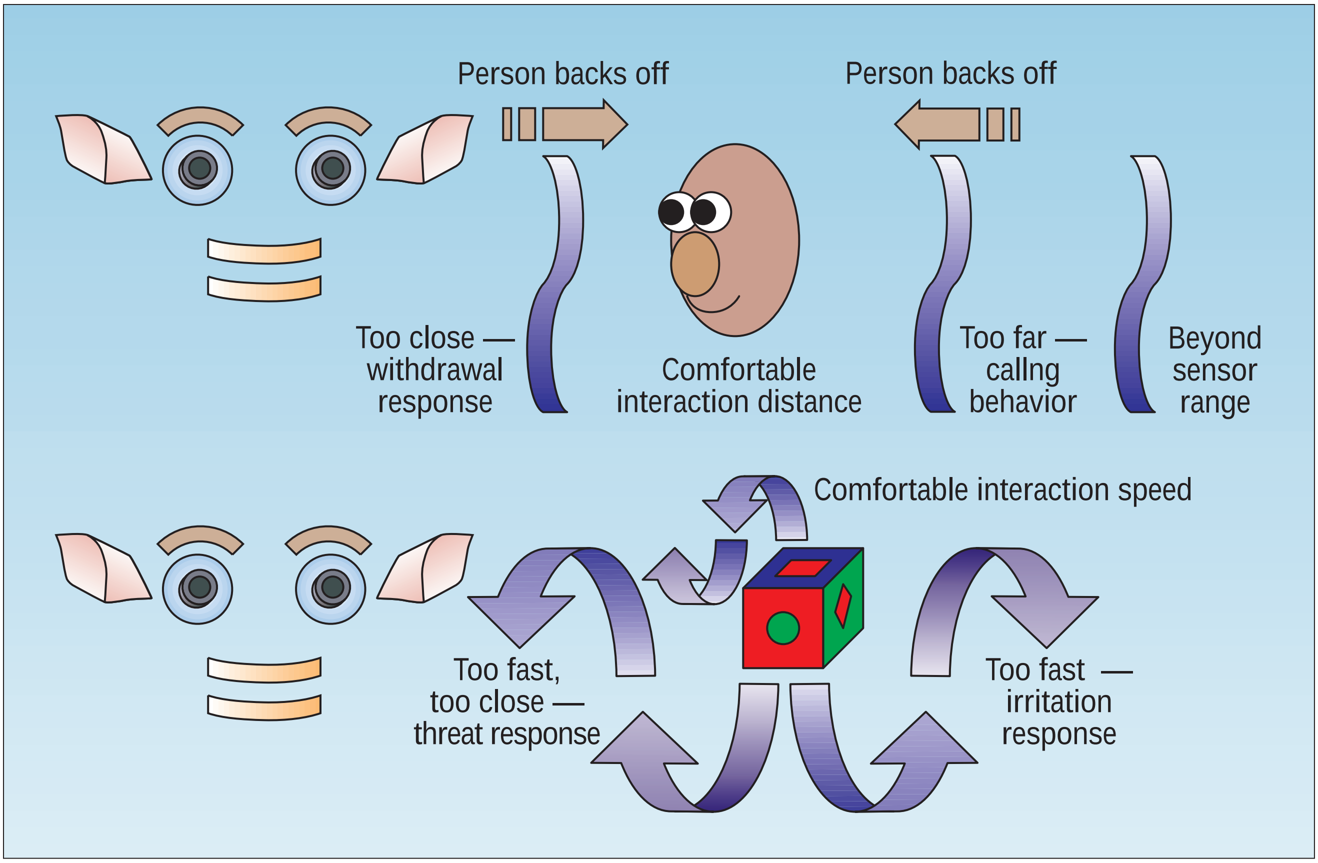
Person backs off
Person backs off
Too close — withdrawal response
Comfortable interaction distance
Too far — callng behavior
Beyond sensor range
Comfortable interaction speed
Too fast, too close — threat response
Too fast — irritation response

Figure 8. Regulating interaction. The robot attracts people too distant to be seen clearly. If they come too close, it signals discomfort and withdraws. The withdrawal moves the robot back physically, but is more effective in signaling to the human to back off. Toys or people that move too rapidly cause irritation.

correct for drift. These two mechanisms work together to give humans stable gaze as the head moves.

We implemented an approximation to the human ocular-motor system for Kismet. Kismet's eyes periodically saccade to new targets chosen by the attention system, tracking them smoothly if they move and the robot wishes to engage them.

Vergence eye movements are more challenging, because disjunctive-eye-movement errors can give the eyes the disturbing appearance of moving independently. Conjunctive-movement errors have a much smaller impact on an observer, because the eyes clearly move in lockstep. We have developed a vestibular-ocular reflex analogue using a 3-axis inertial sensor, and a crude approximation of the opto-kinetic reflex combined with our smooth-pursuit implementation.

Communicative motor acts

Eye movements have communicative value. In addition to indicating the robot's locus of attention, they convey its degree of engagement, communicating how strongly the robot's behavior is organized around what it's looking at. The robot's eyes flicking from place to place without resting indicates a low engagement level, appropriate to a visual search behavior. Prolonged fixation with smooth pursuit and head orientation towards the target conveys a much greater level of engagement, suggesting that the robot's behavior is very strongly focused on that target.

Eye movements are the most obvious and direct motor actions that support visual perception, however postural shifts and fixed action patterns involving the entire robot also have an important role. Kismet has several coordinated motor actions designed to deal with its perceptual limitations (see Figure 8). For example, if a person is visible but too distant for Kismet to view their face at adequate resolution, it engages in a calling behavior to summon the person closer. People who come too close to the robot also cause difficulties for the cameras with narrow fields of view, because only a small part of a face might be visible. In this circumstance, a withdrawal response results, where Kismet draws back physically from the person. This behavior aids the cameras somewhat by increasing the distance between Kismet and the human. But the behavior can have a secondary and greater effect through social amplification—for a human close to Kismet a withdrawal response is a strong social cue to back away, because it is analogous to the human response to personal-space invasions.

We can use similar behavioral patterns to support the visual perception of objects. If an object is too close, Kismet can lean away from it; if it is too far away, Kismet can crane its neck toward it. Again, in a social context, such actions have power beyond their immediate physical consequences. A human, reading intent into the robot's actions, may amplify those actions. For example, they might interpret neck-craning toward a toy as interest in that toy, resulting in the human bringing the toy closer to the robot.

Another limitation of the visual system is how quickly it can track moving objects. If objects or people move at excessive speeds, Kismet has difficulty tracking them continuously. To discourage excessive boisterousness in the movement of people and the objects they manipulate, Kismet shows irritation when its tracker approaches its limits. These limits are either physical (the maximum rate at which the eyes and neck move), or computational (the maximum displacement per frame from the cameras with which the robot searches for a target).

These regulatory mechanisms play roles in more complex social interactions, such as conversational turn-taking. Here, gaze-direction control is important for regulating conversation rate.[10] Generally, people are likely to glance aside when they begin their turn, and make eye contact when they are prepared to relinquish their turn and await a response. People blink most frequently when they end an utterance. These and other cues let Kismet influence the conversation flow to the advantage of its auditory-processing. Here, we see the visual-motor system serve a nominally unrelated sensory modality, just as that system in turn may recruit behaviors only indirectly connected to vision (such as ear wiggling to call someone closer).

These mechanisms also help protect the robot. Objects that suddenly appear close to the robot trigger a looming reflex, causing the robot to quickly withdraw and appear startled. If the event is repeated, the response quickly habituates and the robot simply appears annoyed, because its best strategy for ending these repetitions is to clearly signal their undesirability. Similarly, rapidly moving objects close to the robot are threatening and therefore trigger an escape response.

These mechanisms help elicit natural and intuitive responses from people. But even without them, it is often clear when Kismet's perception is failing, and what corrective action would help, because the robot's behavior reflects its perception in a familiar way. Inferences that we make based on our human preconceptions are actually likely to work.

Motor control for a social robot poses challenges beyond stability and accuracy. Human observers will perceive motor actions as semantically rich, regardless of whether the robot intends the imputed meaning. Such perception, which constrains the robot's physical appearance and movement, can facilitate natural interactions between robot and human. It allows the robot to be readable—to make its behavioral intent and motivational state transparent at an intuitive level to those with whom it interacts. It allows the robot to regulate its interactions to suit its perceptual and motor capabilities in an intuitive way—one with which humans naturally cooperate. And it gives the robot leverage over the world far beyond its physical competence. If properly designed, the robot's visual behaviors can match human expectations and allow both robot and human to participate in natural and intuitive social interactions.

References

1. C. Breazeal and B. Scassellati, "How to Build Robots that Make Friends and Influence People," *Proc. Int'l Conf. Intelligent Robots and Systems*, Kyongju, Korea, 1999.

2. D. Ballard, "Behavioral Constraints on Animate Vision," *Image and Vision Computing*, Vol. 7, No. 1, 1989, pp. 3–9.

3. R.A. Brooks et al., "The Cog Project: Building a Humanoid Robot," *Computation for Metaphors, Analogy and Agents*, C. Nehaniv, ed., *Springer Lecture Notes in Artificial Intelligence*, Vol. 1562, Springer-Verlag, 1998, pp. 52–87.

4. A. Bernardino and J. Santos-Victor, "Binocular Visual Tracking: Integration of Perception and Control," *IEEE Trans. Robotics and Automation*, Vol. 15, No. 6, Dec. 1999, pp. 1080–1094.

5. H.C. Nothdurft, "The Role of Features in Pre-Attentive Vision: Comparison of Orientation, Motion and Color Cues," *Vision Research*, Vol. 33, No. 14, 1993, pp. 1937–1958.

6. C. Breazeal and B. Scassellati, "A Context-Dependent Attention System for a Social Robot," *Proc. 16th Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 1999, Stockholm, pp. 1146–1151.

7. J.M. Wolfe, "Guided Search 2.0: A Revised Model of Visual Search," *Psychonomic Bull. & Rev.*, Vol. 1, No. 2, 1994, pp. 202–238.

8. L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11, 1998, pp. 1254–1259.

9. E.R. Kandel, J.H. Schwarz, and T.M. Jessel, *Principles of Neural Science, 4th Edition*, McGraw-Hill, New York, 2000.

10. J. Cassell, "Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents," *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost and E. Churchill, eds., MIT Press, Cambridge, Mass., 2000.

Cynthia Breazeal received a ScD from MIT in the Department of Electrical Engineering and Computer Science, with Rodney Brooks at the MIT Artificial Intelligence Laboratory. Her interests focus on humanlike robots that can interact in natural, social ways with humans. She received her BS in electrical and computer engineering from the University of Calif., Santa Barbara, and her MS in electrical engineering and computer science from MIT. Contact her at the MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139; cynthia@ai.mit.edu.

Aaron Edsinger received a BS in computer systems at Stanford, and is currently a graduate student with Rodney Brooks at the MIT Artificial Intelligence Laboratory. Contact him at MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139; edsinger@ai.mit.edu.

Paul Fitzpatrick received a BE and ME in computer engineering from the University of Limerick, Ireland, and is currently a graduate student with Rodney Brooks at the MIT Artificial Intelligence Laboratory. Contact him at MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139; paulfitz@ai.mit.edu.

Brian Scassellati is completing his PhD with Rodney Brooks at the MIT Artificial Intelligence Laboratory. His work is strongly grounded in theories of how the human mind develops, and he is interested in robotics as a tool for evaluating models from biological sciences. He received a BS in computer science, a BS in brain and cognitive science, and his ME in electrical engineering and computer science from MIT. Contact him at the MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139; scaz@ai.mit.edu.

Paulina Varchavskaia received a BSc in computer science with cognitive science from University College London, and is currently a graduate student with Rodney Brooks at the MIT Artificial Intelligence Laboratory. Contact her at MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139; paulina@ai.mit.edu.