Social Constraints on Animate Vision

Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, Brian Scassellati, Paulina Varchavskaia

MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139
USA
{cynthia,edsinger,paulfitz,scaz,paulina}@ai.mit.edu

Abstract

Our group builds robots to operate in natural, social environments. The challenge of interacting with humans constrains how our robots appear physically, how they move, how they perceive the world, and how their behaviors are organized. This article describes an integrated visual-motor system we have constructed that negotiates between the physical constraints of the robot, the perceptual needs of the robot's behavioral and motivational systems, and the social implications of motor acts.

Keywords

Robotics, humanoid, active vision, social, attention, regulation

IEEE Intelligent Systems

Abstract

Our group builds robots to operate in natural, social environments.[1] The challenge of interacting with humans constrains how our robots appear physically, how they move, how they perceive the world, and how their behaviors are organized. This article describes an integrated visual-motor system we have constructed that negotiates between the physical constraints of the robot, the perceptual needs of the robot's behavioral and motivational systems, and the social implications of motor acts.

Introduction

For robots and humans to interact meaningfully, it is important that they understand each other enough to be able to shape each other's behavior. This has several implications. One of the most basic is that robot and human should have at least some overlapping perceptual abilities. Otherwise, they can have little idea of what the other is sensing and responding to. Vision is one important sensory modality for human interaction, and the one we focus on in this article. We endow our robots with visual perception that is human-like in its physical implementation.

Similarity of perception requires more than similarity of sensors. Not all sensed stimuli are equally behaviorally relevant. It is important that both human and robot find the same types of stimuli salient in similar conditions. Our robots have a set of perceptual biases based on the human pre-attentive visual system. These biases can be modulated by the motivational state of the robot, making later perceptual stages more behaviorally relevant. This approximates the top-down influence of motivation on the bottom-up pre-attentive process found in human vision.

Visual perception requires high bandwidth and is computationally demanding. In the early stages of human vision, the entire visual field is processed in parallel. Later computational steps are applied much more selectively, so that behaviorally relevant parts of the visual field can be processed in greater detail. This mechanism of visual attention is just as important for robots as it is for humans, from the same considerations of resource allocation. The existence of visual attention is also key to satisfying the expectations of humans concerning what can and cannot be perceived visually. We have implemented a context-dependent attention system that goes some way towards this.

Human eye movements have a high communicative value. For example, gaze direction is a good indicator of the locus of visual attention. Knowing a person's locus of attention reveals what that person currently considers behaviorally relevant, which is in turn a powerful clue to their intent. The dynamic aspects of eye movement, such as staring versus glancing, also convey information. Eye movements are particularly potent during social interactions, such as conversational turn-taking, where making and breaking eye contact plays an important role in regulating the exchange. We model the eye movements of our robots after humans, so that they may have similar communicative value.

Our hope is that by following the example of the human visual system, the robot's behavior will be easily understood because it is analogous to the behavior of a human in similar circumstances (see Figure 1). For example, when an anthropomorphic robot moves its eyes and neck to orient toward an object, an observer can effortlessly conclude that the robot has become interested in that object. These traits lead not only to behavior that is easy to understand but also allows the robot's behavior to fit into the social norms that the person expects.

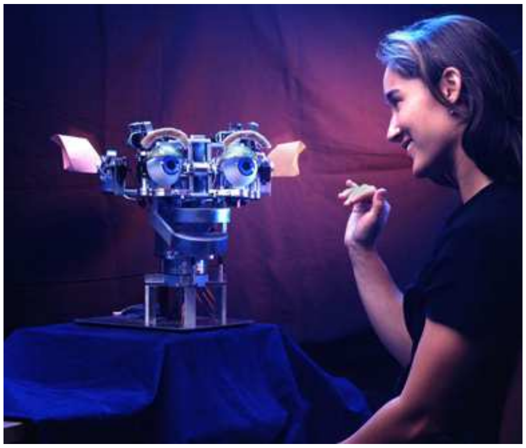


Figure 1: Kismet, a robot capable of conveying intentionality through facial expressions and behavior.[2] Here, the robot's physical state expresses attention to and interest in the human beside it. Another person – for example, the photographer – would expect to have to attract the robot's attention before being able to influence its behavior.

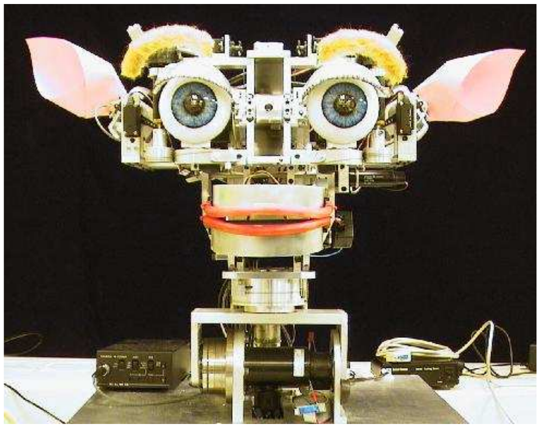
There are other advantages to modeling our implementation after the human visual system. There is a wealth of data and proposed models for how the human visual system is organized. This data provides not only a modular decomposition but also mechanisms for evaluating the performance of the complete system.

Another advantage is robustness. A system that integrates action, perception, attention, and other cognitive capabilities can be more flexible and reliable than a system that focuses on only one of these aspects. Adding additional perceptual capabilities and additional constraints between behavioral and perceptual modules can increase the relevance of behaviors while limiting the computational requirements.[3] For example, in isolation, two difficult problems for a visual tracking system are knowing what to track and knowing when to switch to a new target. These problems can be simplified by combining the tracker with a visual attention system that can identify objects that are behaviorally relevant and worth tracking. In addition, the tracking system benefits the attention system by maintaining the object of interest in the center of the visual field. This simplifies the computation necessary to implement behavioral habituation. These two modules work in concert to compensate for the deficiencies of the other and to limit the required computation in each.

Physical form

Currently, the most sophisticated of our robots in terms of visual-motor behavior is Kismet. This robot is an active vision head augmented with expressive facial features (see Figure 2). Kismet is designed 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. Kismet has three degrees of freedom to control gaze direction, three degrees of freedom to control its neck, and fifteen degrees of freedom in other expressive components of the face (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, there are two unobtrusive central cameras fixed with respect to the head, each with a wider field of view but correspondingly lower acuity. The reason for this mixture of cameras is that typical visual tasks require both high acuity and a wide field of view. High acuity is needed for recognition tasks and for controlling precise visually guided motor movements. A wide field of view is needed for search tasks, for tracking multiple objects, compensating for involuntary ego-motion, etc. A common trade-off found in biological systems is to sample part of the visual field at a high enough resolution to support the first set of tasks, and to sample the rest of the field at an adequate level to support the second set. This is seen in animals with foveate vision, such as humans, where the density of photoreceptors is highest at the center and falls off dramatically towards the periphery. This can be implemented by using specially designed imaging hardware, space-variant image sampling[4], or by using multiple cameras with different fields of view, as we have done.



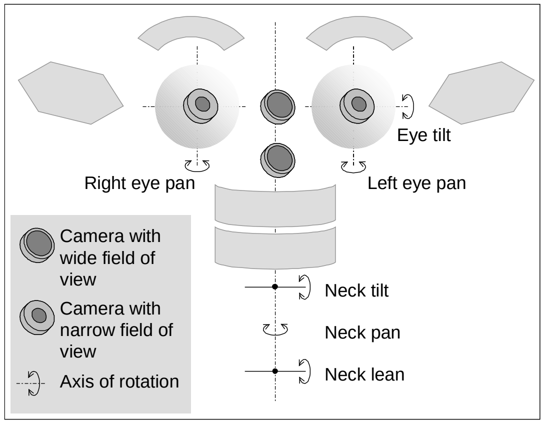


Figure 2: Kismet has a large set of expressive features – eyelids, eyebrows, ears, jaw, lips, neck and eye orientation. The schematic on the right shows the degrees of freedom relevant to visual perception (omitting the eyelids!). The eyes can turn independently along the horizontal (pan), but turn together along the vertical (tilt). 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. These cameras are unaffected by the orientation of the eyes.

Another of our robots, Cog, follows the human sensing arrangement more closely than does Kismet. Cog is a 22 degree of freedom upper-torso humanoid. The mechanical design of the head and neck are based on human anatomy and performance. Each of Cog's eyes has two color CCD cameras, one with a wide field of view for peripheral vision and one with a narrow field of view for high acuity vision – as opposed to Kismet's arrangement, where the wide cameras are fixed with respect to the head. Cog also has a three-axis inertial package that detects head rotation and a gravity vector similar to the human vestibular system.

The designs of our robots are constantly evolving. New degrees of freedom are added, old degrees of freedom are reorganized, sensors are replaced or rearranged, new sensory modalities are introduced. The descriptions given here should be treated as a fleeting snapshot of the current state of the robots.

System architecture

Our hardware and software control architectures have been designed to meet the challenge of real-time processing of visual signals (approaching 30 Hz) with minimal latencies. Kismet's vision system is implemented on a network of nine 400 MHz commercial PCs running the QNX real-time operating system (see Figure 3). Kismet's motivational system runs on a collection of four Motorola 68332 processors. Machines running Windows NT and Linux are also networked 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 on line.

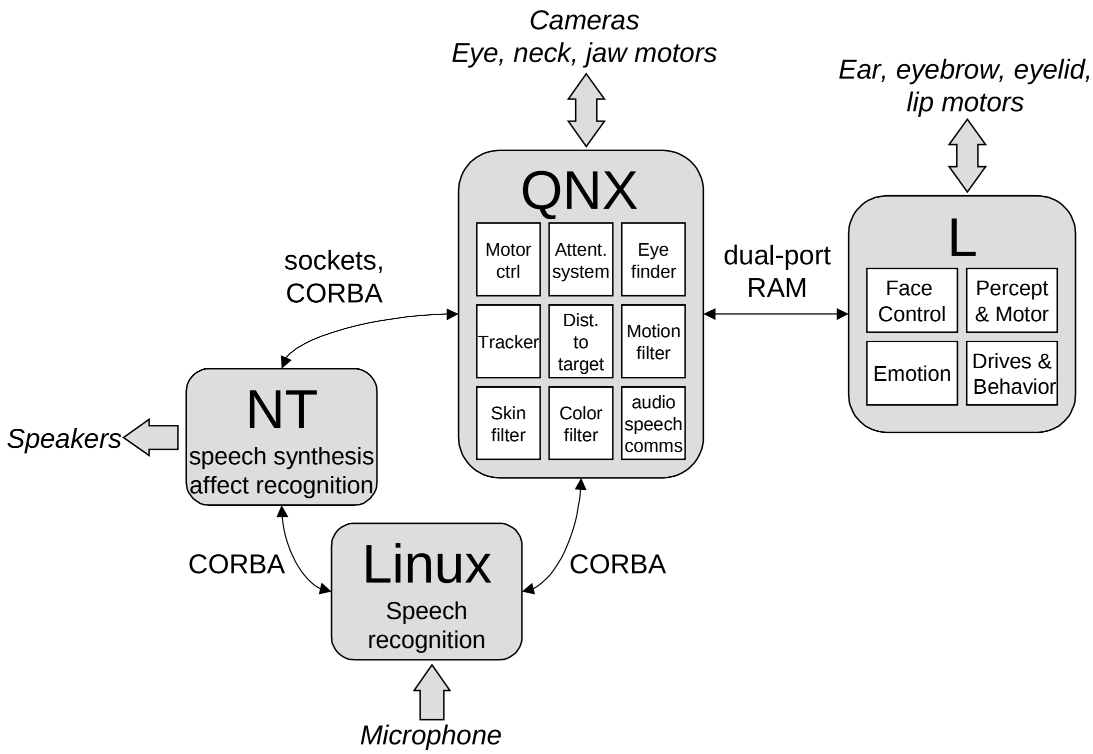


Figure 3: System architecture for Kismet. The motivation system runs on four Motorola 68332 microprocessors running L, a multi-threaded Lisp developed in our lab. Vision processing and eye/neck control is performed by nine networked PCs running QNX.

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 a variety of 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. Examples of features we have used are shown in Figure 4. Looming objects are also detected preattentively, to facilitate a fast reflexive withdrawal.

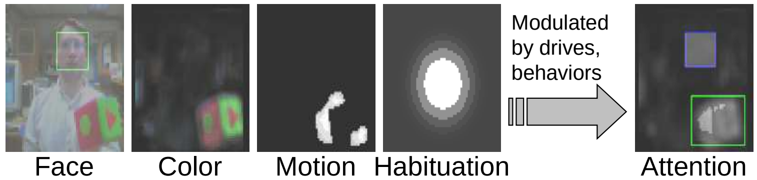


Figure 4: The robot's attention is determined by a combination of low-level perceptual stimuli. The relative weightings of the stimuli are modulated by high-level behavior and motivational influences.[6] A sufficiently salient stimulus in any modality can pre-empt attention, similar to the human response to sudden motion. All else being equal, larger objects are considered more salient than smaller ones. The design is intended to keep the robot responsive to unexpected events, while avoiding making it a slave to every whim of its environment. With this model, people intuitively provide the right cues to direct the robot's attention (shake object, move closer, wave hand, etc.).

Visual attention

We have implemented Wolfe's model of human 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. A variety of low-level feature detectors (such as color, motion, and shape) are combined through a weighted average to produce a single attention map. This combination allows the robot to select regions that are visually salient and to direct its computational and behavioral resources towards 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.

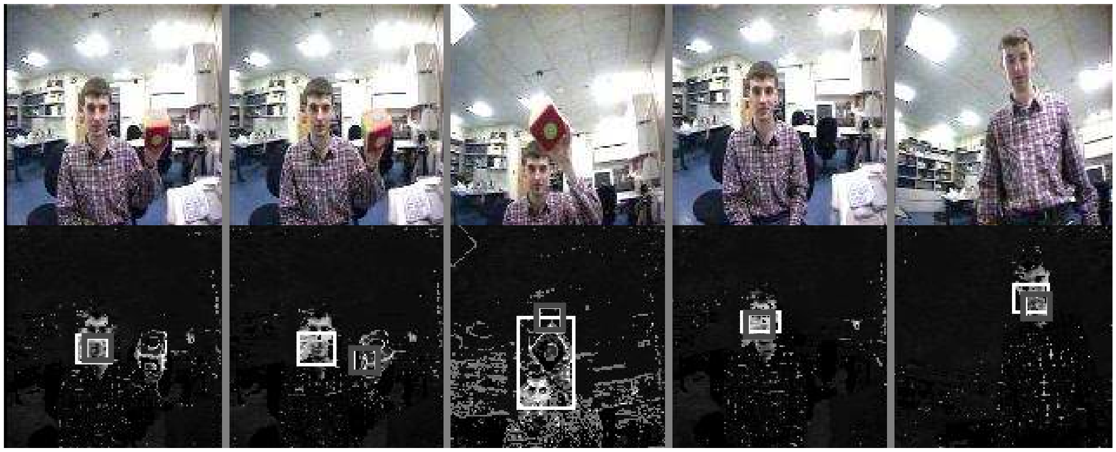


Figure 5: Manipulating the robot's attention. Images on the top row are from Kismet's upper wide camera. Images on the bottom summarize the contemporaneous state of the robot's attention system. Brightness in the lower image corresponds to salience; rectangles correspond to regions of interest. The thickest 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 locus of 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, which is tracked as it moves.

Post-attentive processing

Once the attention system has selected regions of the visual field that are potentially behaviorally relevant, more intensive computation can be applied to these regions than could be applied across the whole field. 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 gaze to a locus of attention, so that a relatively high resolution image of the area being searched is available from the narrow field of view cameras (Figure 6). Another calculation currently done post-attentively is distance to a target. This distance is estimated using a stereo match between the two central cameras.

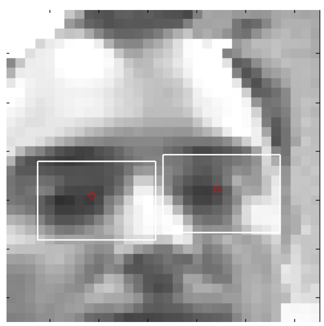


Figure 6: Eyes are searched for within a restricted part of the robot's field of view.

Eye movement primitives

Kismet's visual-motor control is modeled after the human ocular-motor system. The human system is so good at providing a stable percept of the world that we have no intuitive appreciation of the physical constraints under which it operates.

Humans have foveate vision. The fovea (the center of the retina) has a much higher density of photoreceptors than the periphery. This means that to see an object clearly, humans must move their eyes such that the image of the object falls on the fovea. Human eye movement is not smooth. It is composed of 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 period of fixation, during which the eyes are relatively stable. They are by no means stationary, and continue to engage in corrective micro-saccades and other small movements. If the eyes fixate on a moving object, they can follow it with a continuous tracking movement called smooth pursuit. This type of eye movement cannot be evoked voluntarily, but only occurs in the presence of a moving object. Periods of fixation typically end after some hundreds of milliseconds, after which a new saccade will occur.[9]

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

Since 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 orientation of the eyes stable as the eyes move. This is a very fast response, but is prone to the accumulation of error over time. The opto-kinetic response is a slower compensation mechanism that uses a measure of the visual slip of the image across the retina to correct for drift. These two mechanisms work together to give humans stable gaze as the head moves.

Our implementation of an ocular-motor system is an approximation of the human system. 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, since errors in disjunctive eye movements can give the eyes a disturbing appearance of moving independently. Errors in conjunctive movements have a much smaller impact on an observer, since the eyes clearly move in lock-step. An analogue of the vestibular-ocular reflex has been developed for Cog using a 3-axis inertial sensor. A crude approximation of the opto-kinetic reflex is rolled into our implementation of smooth pursuit.

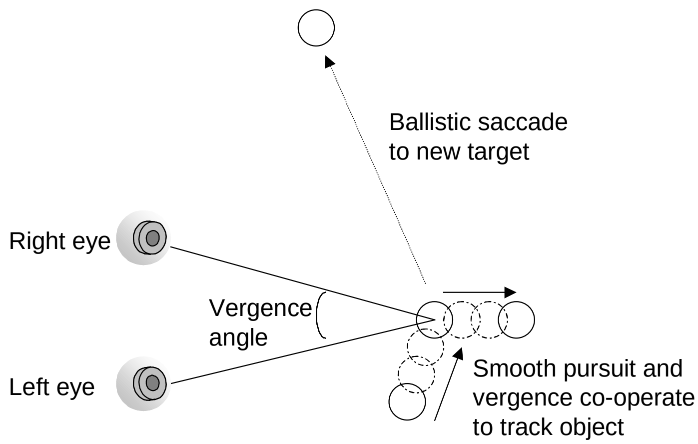


Figure 7: Humans exhibit four characteristic types of eye motion. Saccadic movements are high-speed ballistic motions that center a target in the field of view. Smooth pursuit movements are used to track a moving object at low velocities. The vestibulo-ocular and opto-kinetic reflexes act to maintain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

Communicative motor acts

Eye movements have communicative value. As discussed previously, they indicate the robot's locus of attention. The robot's degree of engagement can also be conveyed, to communicate how strongly the robot's behavior is organized around what it is currently looking at. If the robot's eyes flick about from place to place without resting, that indicates a low level of engagement, appropriate to a visual search behavior. Prolonged fixation with smooth pursuit and orientation of the head towards the target conveys a much greater level of engagement, suggesting that the robot's behavior is very strongly organized about the locus of attention.

Eye movements are the most obvious and direct motor actions that support visual perception. But they are by no means the only ones. Postural shifts and fixed action patterns involving the entire robot also have an important role. Kismet has a number of coordinated motor actions designed to deal with various limitations of Kismet's visual perception (see Figure 8). For example, if a person is visible, but is too distant for their face to be imaged at adequate resolution, Kismet 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, since only a small part of a face may be visible. In this circumstance, a withdrawal response is invoked, where Kismet draws back physically from the person. This behavior, by itself, 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, since it is analogous to the human response to invasions of "personal space."

Similar kinds of behavior can be used 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 towards 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, neck-craning towards a toy may be interpreted 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 bias people away from excessively boisterous behavior in their own movements or in the movement of objects they manipulate, Kismet shows irritation when its tracker is at the limits of its ability. 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 over which a target is searched for).

Such regulatory mechanisms play roles in more complex social interactions, such as conversational turn-taking. Here control of gaze direction is important for regulating conversation rate.[10] In general, 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. Blinks occur most frequently at the end of an utterance. These and other cues allow Kismet to influence the flow of conversation to the advantage of its auditory processing. Here we see the visual-motor system being driven by the requirements of a nominally unrelated sensory modality, just as behaviors that seem completely orthogonal to vision (such as ear-wiggling during the call behavior to attract a person's attention) are nevertheless recruited for the purposes of regulation.

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, since its best strategy for ending these repetitions is to clearly signal that they are undesirable. Similarly, rapidly moving objects close to the robot are threatening and trigger an escape response.

These mechanisms are all designed to elicit natural and intuitive responses from humans, without any special training. But even without these carefully crafted mechanisms, it is often clear to a human when Kismet's perception is failing, and what corrective action would help, because the robot's perception is reflected in behavior in a familiar way. Inferences made based on our human preconceptions are actually likely to work.

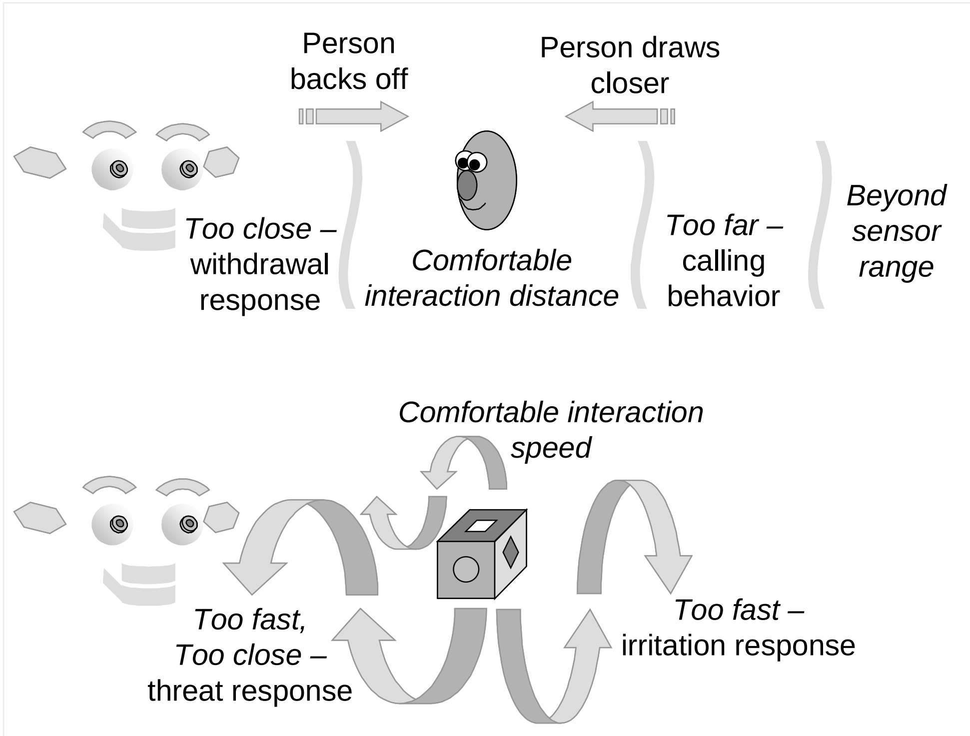


Figure 8: Regulating interaction. People too distant to be seen clearly are called closer; if they come too close, the robot signals discomfort and withdraws. The withdrawal moves the robot back somewhat physically, but is more effective in signaling to the human to back off. Toys or people that move too rapidly cause irritation.

Conclusions

Motor control for a social robot poses challenges beyond issues of stability and accuracy. Motor actions will be perceived by human observers as semantically rich, regardless of whether the imputed meaning is intended or not. This can be a powerful resource for facilitating natural interactions between robot and human, and places constraints on the robot's physical appearance and movement. It allows the robot to be readable – to make its behavioral intent and motivational state transparent at an intuitive level to those it interacts with. It allows the robot to regulate its interactions to suit its perceptual and motor capabilities, again in an intuitive way with which humans naturally co-operate. And it gives the robot leverage over the world that extends far beyond its physical competence, through social amplification of its perceived intent. If properly designed, the robot's visual behaviors can be matched to human expectations and allow both robot and human to participate in natural and intuitive social interactions.

References

1. R. A. Brooks, C. Breazeal, R. Irie, C. C. Kemp, M. J. Marjanovic, B. Scassellati and M. M. Williamson. *The Cog Project: Building a Humanoid Robot,* in C. Nehaniv, ed., Computation for Metaphors, Analogy and Agents, Vol. 1562 of Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998.

2. C. Breazeal and B. Scassellati. *How to Build Robots that Make Friends and Influence People,* Proceedings of the International Conference on Intelligent Robots and Systems, Kyongju, Korea, 1999.

3. D. Ballard. *Behavioral Constraints on Animate Vision*, Image and Vision Computing, 7(1):3-9, 1989.

4. A. Bernardino and J. Santos-Victor. *Binocular Visual Tracking: Integration of Perception and Control*, IEEE Transactions on Robotics and Automation, (15)6, Dec. 1999.

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

6. C. Breazeal and B. Scassellati. *A context-dependent attention system for a social robot*. IJCAI 1999.

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

8. L. Itti, C. Koch and E. Niebur. *A model of saliency-based visual attention for rapid scene analysis*, IEEE

Transactions on Pattarn Analysis and Machine Intelligence, 20(11):1254-1259, 1998.

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

10. J. Cassell. *Embodied conversation: integrating face and gesture into automatic spoken dialogue systems.* Luperfoy (ed.) Spoken Dialogue Systems, MIT Press (to appear).

Biographical Sketches

Cynthia Breazeal received her B.Sc. degree from the University of California, Santa Barbara in Electrical and Computer Engineering in 1989, and received her M.Sc. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 1993 She is currently completing her Ph.D. with Prof. Rodney Brooks at the MIT Artificial Intelligence Laboratory. Her current interests focus on human-like robots that can interact in natural, social ways with humans.

Aaron Edsinger received a B.S. in Computer System at Stanford, and is currently a graduate student with Prof. Rodney Brooks at the MIT Artificial Intelligence Laboratory.

Paul Fitzpatrick received a B.Eng and M.Eng. in Computer Engineering at the University of Limerick, Ireland, and is currently a graduate student with Prof. Rodney Brooks at the MIT Artificial Intelligence Laboratory.

Brian Scassellati received S.B. degrees in computer science and brain and cognitive science from the Massachusetts Institute of Technology in 1994, and a Masters of Engineering degree in Electrical Engineering and Computer Science from MIT in 1995. Since then, he has been a graduate student working towards his Ph.D. with Prof. 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 utilizing robotics as a tool for evaluating models from biological sciences.

Paulina Varchavskaia received a B.Sc. in Computer Science with Cognitive Science at University College London, and is currently a graduate student with Prof. Rodney Brooks at the MIT Artificial Intelligence Laboratory.