That Certain Look: Social Amplification of Animate Vision

Cynthia Breazeal and Paul Fitzpatrick
MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139
USA
{cynthia,paulfitz}@ai.mit.edu

Abstract

Animate vision control for a social robot poses challenges beyond issues of stability and accuracy, as well as advantages beyond computational efficiency and perceptual robustness[1]. We have found that the human-like eye movements of a robot have high communicative value to the humans that interact with it. This can be a powerful resource for facilitating natural interactions between robot and human. 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. This paper describes a variety of ways we are exploring and exploiting the communicative value of robotic gaze, in concert with facial display and body posture, to improve the quality of interaction between robot and human.

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 focused on in this paper. We endow our robots with visual perception that is human-like in its physical implementation [2].

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 [3]. 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.

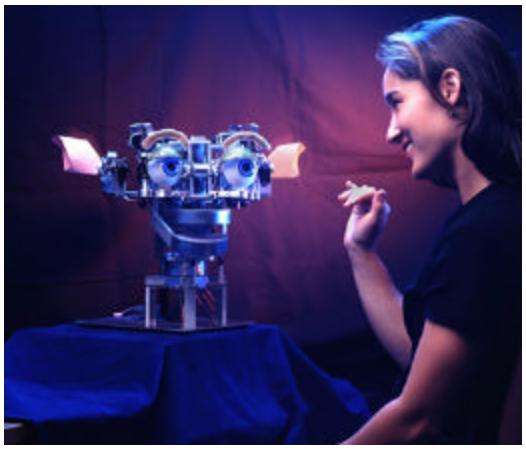

Figure 1: Kismet, a robot capable of conveying intentionality through facial expressions and behavior [4]. Here, the robot's physical state expresses attention to and interest in the human beside it. Another person – for example, the photographer – would expect to have to attract the robot's attention before being able to influence its behavior.

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, or by using multiple cameras with different fields of view, as we have done.

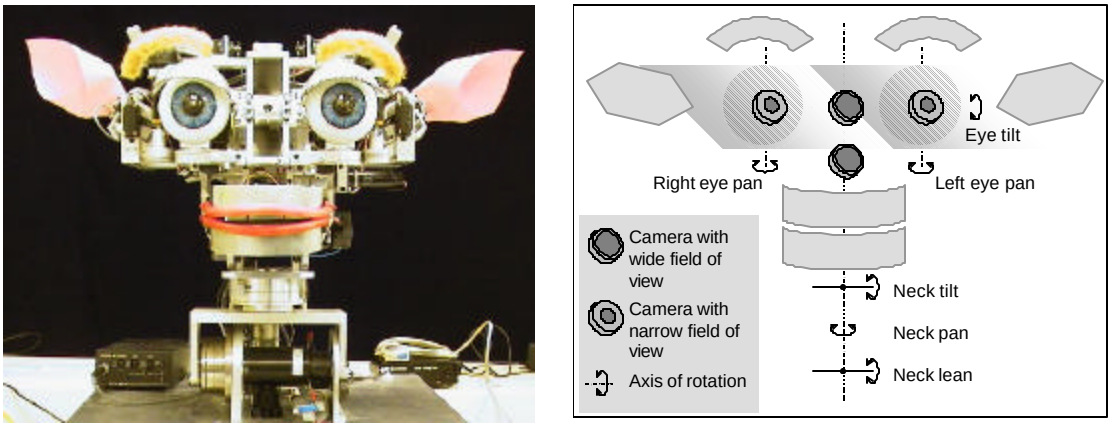

Figure 2: Kismet has a large set of expressive features – eyelids, eyebrows, ears, jaw, lips, neck and eye orientation. The schematic on the right shows the degrees of freedom relevant to visual perception (omitting the eyelids!). The eyes can turn independently along the horizontal (pan), but turn together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

Communicative motor acts

In previous work, we have reported on Kismet's visual capabilities [5]. These include low level feature extractors for color, skin tone, and motion which are combined by a context-dependent attention system that enables the robot to direct its attention to the most perceptually and behaviorally relevant. Post attentive visual processing includes finding eyes on skin-toned areas and computing the distance to the most salient target. The robot also has a repertoire of active vision behaviors including saccades, smooth pursuit, a simple opto-kinetic response, and neck/eye orientation movements. We have argued the benefits of this system in terms of robustness and computational efficiency.

However, Kismet interacts with people and its eye movements also have communicative value to the human who interacts with it. 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 3). 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 [3] 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.

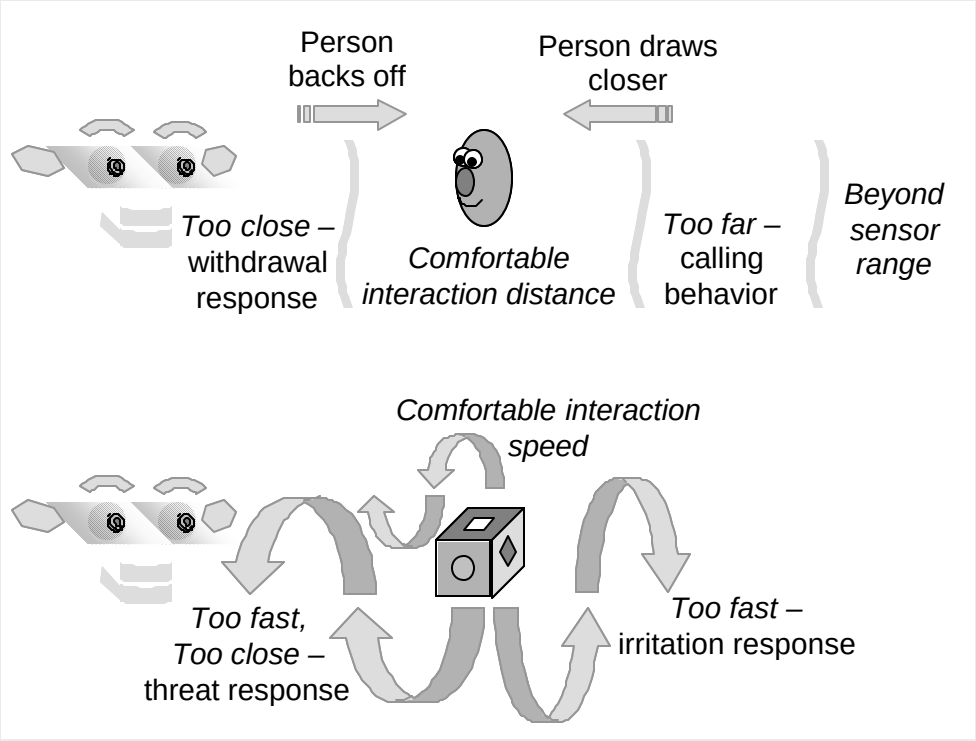

Figure 3: Regulating interaction. People too distant to be seen clearly are called closer; if they come too close, the robot signals discomfort and withdraws. The withdrawal moves the robot back somewhat physically, but is more effective in signaling to the human to back off. Toys or people that move too rapidly cause irritation.

We have begin a series of studies with naïve subjects to observe how they interpret and respond to Kismet's cues. Thus far, they report that the eyes are its most compelling feature. When trying to attract the robot's attention, the subjects use the robot's gaze direction as the critical cue. Furthermore, it is not sufficient that the robot merely orients to the object but must also smoothly pursue it before the subject indicates that the robot's attention has been acquired. When subjects come very close to the robot, causing it to withdraw with a stern expression, they recognize that they have invaded Kismet's "personal space" and tend to back away. They report that maintaining eye contact through vocal exchanges is important. They also report the raising of the robot's eyebrows as an important cue for taking their turn in the exchange. Studies are continuing, but these early findings seem to support the social amplification process.

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. D. Ballard. *Behavioral Constraints on Animate Vision,* Image and Vision Computing, 7(1):3-9, 1989.
2. E. R. Kandel, J. H. Schwarz and T. M. Jessel. *Principles of Neural Science,* 4th Edition, McGraw-Hill, 2000.
3. J. Cassell. *Embodied conversation: integrating face and gesture into automatic spoken dialogue systems.* Luperfoy (ed.) Spoken Dialogue Systems, MIT Press (to appear).
4. 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.
5. C. Breazeal Aaron Edsinger, Paul Fitzpatrick, Brian Scassellati, Paulina Varchavskaia. *Social Constraints on Animate Vision*. IEEE 2000 special issue on humanoid robotics, forthcoming.