Saliency in Human-Computer Interaction *

Polly K. Pook
MIT AI Lab
545 Technology Square
Cambridge, MA 02139 USA
pook@ai.mit.edu

Abstract

This paper considers ways in which a person can cue and constrain an artificial agent's attention to salient features. In one experiment, a person uses gestures to direct an otherwise autonomous robot hand through a known task. Each gesture instantiates the key spatial and intentional features for the task at that moment in time. In a second experiment, which is work in progress, a person will use speech and gesture to assist an "intelligent room" in learning to recognize the objects in its environment. In this case, the robot (the room) will take both direction and correction signals from the person and use them to tune its feature saliency map and limit its search space.

1 Introduction

The workshop precis asks, how can robotic technology assist people? This paper considers an opposing question, how can people assist robots? The topics are complementary in a least one way: both benefit by reducing the human-computer interface to its essentials.

A person rarely wants explicit control over all motor actions of the assistive agent, be it a wheelchair or a manipulator. Indeed the person may be unable to finely control the agent, regardless of volition. Instead, the person should have strategic control while the agent governs its own tight-looped servo control.

Similarly, the person who assists an artificial agent naturally limits supervision in both throughput and bandwidth. No person is going to provide feedback on each of 10,000 simulated trials, as is often required by current machine learning techniques such as genetic algorithms and reinforcement learning. Instead, one must ask the person to provide a few key directives that are otherwise beyond the cognitive capabilities of the agent. Thus the overriding goal is to distill human communication to the essential, or salient, features required by the agent.

*The work on *teleassistance* was conducted at the University of Rochester and supported by NSF and the Human Science Frontiers Program. Research in the MIT Intelligent Room is supported by DARPA.

Determining saliency is a significant problem in AI. To us, what's so striking about a visual scene or a physical force is, to an artificial agent, often indistinguishable from the rest of the sensorium. The difficulty is compounded by context dependency: features that predominate in one setting are often irrelevant to another. The agent therefore could benefit greatly from on-line human cues.

This paper considers ways in which a person can cue and constrain an artificial agent's attention to salient features. In one experiment, a person uses gestures to direct an otherwise autonomous robot hand through a known task. Each gesture instantiates the key spatial and intentional features for the task at that moment in time. In a second experiment, which is work in progress, a person will use speech and gesture to assist an "intelligent room" in learning to recognize the objects in its environment. In this case, the robot (the room) will take both direction and correction signals from the person and use them to tune its saliency map and limit its search space.

The work proposed here limits consideration to features that are rather low-level: color, shape, direction, etc. These features are variables in cognition at the embodiment level. The research may have application to areas such as prosthetics, but it is worth noting that assistive technology certainly seeks higher-level control as well.

2 Feature extraction in sequential cognitive operations

Embodiment

To describe phenomena that occur at different time scales, computational models of the brain necessarily must incorporate different levels of abstraction. We have argued that at time scales of approximately one-third of a second, orienting movements of the body play a crucial role in cognition and form a useful computational level [Ballard, Hayhoe, Pook & Rao, 1996]. This level is more abstract than that used to capture neural phenomena yet is framed at a level of abstraction below that traditionally used to study high-level cognitive processes such as reading. We term this level

the embodiment level. At the embodiment level, the constraints of the physical system determine the nature of cognitive operations. In humans, we find evidence that the natural sequentiality of body movements can be matched to the natural computational economies of sequential decision systems, at time scales of about one-third second [Kowler & Anton, 1987] , [Ballard et al., 1992]. This is accomplished through a system of implicit reference termed *deictic*, whereby pointing movements are used to bind objects in the world to cognitive programs.

Deictic reference in humans

Ballard, Hayhoe, and Pelz propose that only those features that are key to the particular cognitive operation are extracted at each deictic binding. They test eye movements in a series of block copying experiments. The subjects are asked to assemble a set of colored blocks to match a pre-assembled configuration. Most subjects look twice at each block to be copied, once before selecting a new block and again before placing the new block in the proper configuration. They posit that the subject perceives the color of the block on the first saccade and the location on the second, rather than extract both features simultaneously. The eye movement is a deictic pointer that binds a cognitive behavior, such as color detection, to a specific target.

Eye fixation is a common deictic device to bind perceptual and motor actions to a specific area or object. Similarly, body position creates a reference for motor actions [Pelz et al., 1994]. Additionally, bi-manual animals often use use one hand as a vise and the other for dexterous manipulation [Guiard, 1987]. The vise hand may mark a reference frame allowing dexterous motions to be made relative to that frame.

Deictic reference in robots

In robotics, deictic variables can define relative coordinate frames for successive motor behaviors [Agre & Chapman 1987]. Such variables can avoid world-centered geometry that varies with robot movement. To open a door, for instance, looking at the doorknob defines a relative servo target. [Crisman and Clearly, 1994] demonstrate the computational advantage of target-centered reference frames for mobile robot navigation. Hand and body position also provide a relative reference frame. Since morphology determines much of how hands are used, the domain knowledge inherent in the shape and frame position can be exploited. The features salient to the task (direction, force) can be extracted and interpreted within the constraints of the reference frame.

3 Example 1: Deictic gestures for robot control

Understanding the deictic references used to bind cognitive programs gives us a starting model for human/robot interaction. In this model, which we call *teleassistance* [Pook and Ballard, 1994] the human provides the deictic references via hand gestures and an otherwise autonomous robot carries out the motor programs. A gesture selects the next motor program to perform and tunes it with hand-centered markers. This illustrates a way of decoupling the human's link between motor program and reflexes. Here the output of the human operator is a deictic code for a motor program that a robot then carries out. This allows the study of the use and necessary properties of the deictic code for situated, autonomous robot action.

The dual-control strategy of teleassistance combines teleoperation and autonomous servo control to their advantage. The use of a simple sign language helps to alleviate many problems inherent to literal master/slave teleoperation. Conversely, the integration of global operator guidance and hand-centered coordinate frames permits the servo routines to position the robot in relative coordinates and perceive features in the feedback within a constrained context, significantly simplifying the computation and reducing the need for detailed task models.

In these experiments the human operator wears a data glove (an EXOS hand master) to communicate the gestures, such as pointing to objects and adopting a grasp preshape. Each sign indicates intention: e.g., reaching or grasping; and, where applicable, a spatial context: e.g., the pointing axis or preshape frame. The robot, a Utah/MIT hand on a Puma arm, acts under local servo control within the proscribed contexts.

Opening a door

The gestural language is very simple. To assist a robot to open a door requires only three signs: point, preshape, and halt. Pointing to the door handle prompts the robot to reach toward it and provides the axis along which to reach. A finite state machine (FSM) for the task specifies the flow of control (Figure 1). This embeds gesture recognition and motor response within the overall task context.

Pointing and preshaping the hand create hand-centered spatial frames. Pointing defines a relative axis for subsequent motion. In the case of preshaping, the relative frame attaches within the opposition space [Arbib et al., 1985] of the robot fingers. For example, a wrap grasp defines a coordinate system relative to the palm. With adequate dexterity and compliance, simply flexing the robot fingers toward the origin of the frame coupled with a force control loop suffices to form a stable grasp. Since the motor action is bound to the local context, the same grasping action can be applied to different objects – a spatula, a mug, a doorknob – by changing the preshape.

The two salient features for each motor program in this task are direction, specified by the hand signs, and force, specified as a significant change in tension in any of the finger joints. Force is perceived identically but

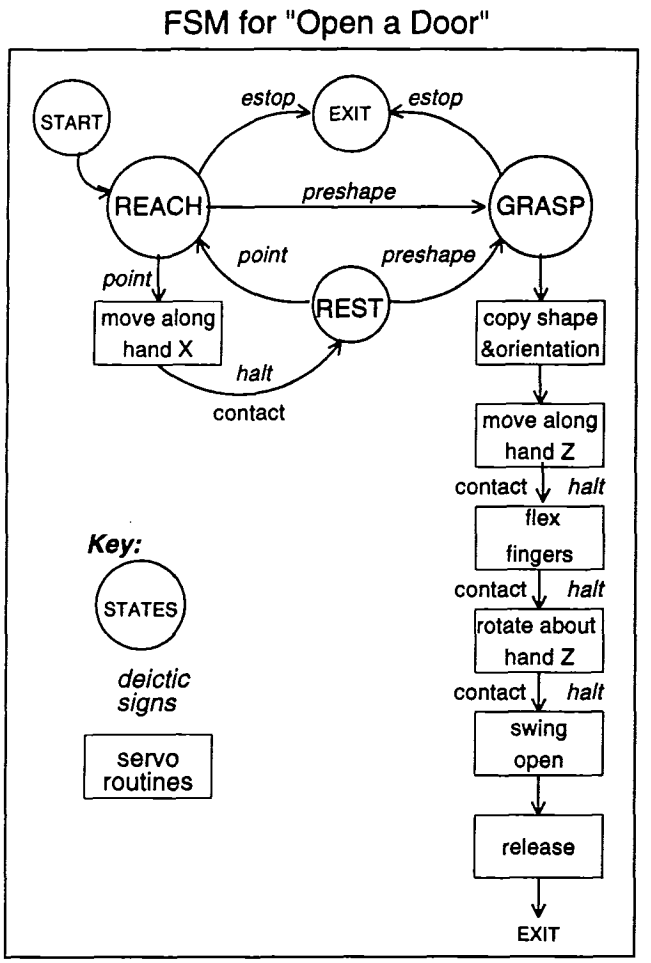

Figure 1: A finite state machine (FSM) for opening a door. The operator's hand sign causes a transition to the appropriate state and, from there, to the corresponding servo routines. The state determines the motor program to execute. The deictic hand signs, *point* and *preshape*, provide a spatial coordinate frame. The routines servo on qualitative changes in joint position error that signify a force contact. Each contact is interpreted within the current context, e.g., bumping the door or reaching the knob's mechanical stop. No special force model is needed. The operator can also push the flow of control through the servo routines by signaling the *halt* hand sign.

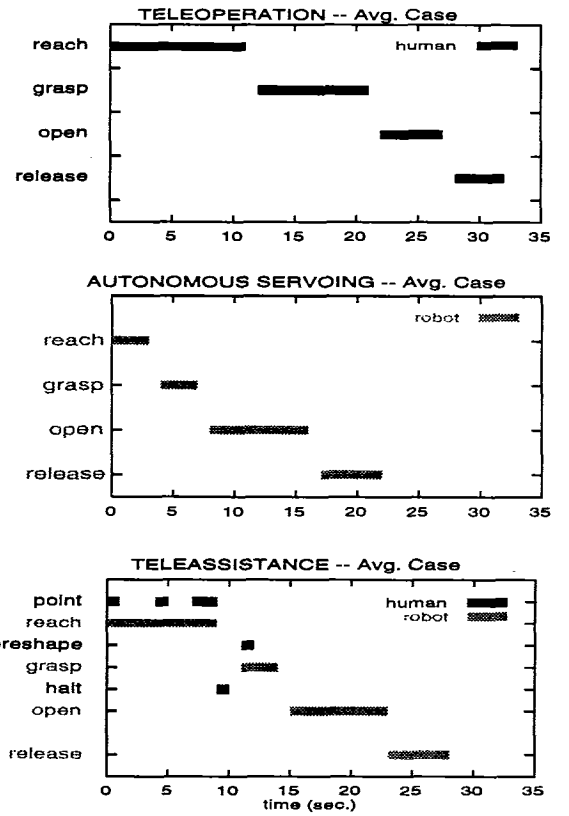

Figure 2: Plots of the time that the human operator (dark bars) and the autonomous robot routines (grey bars) actively control the robot during each phase of the task, under the three control strategies. The teleoperator (top) must supervise 100% of the task; under autonomous control (middle), the robot is fully in charge but with limited strategic abilities; in teleassistance (bottom) the operator supervises the robot only 25% of the time for this task. Once the hand is teleassisted to a position near the door handle, the robot completes the task autonomously.

interpreted differently according to the motor program. While reaching, force suggests that the door has been reached, while turning a force is interpreted as contact with the mechanical stop, i.e., the knob is fully turned. The bound context permits the program to constrain perception to the salient feature and to interpret it in a dynamic way. No special model of the physics of the interaction is needed.

Results

Figure 2 compares robot control by deictic gestures to two other robot control strategies. The first is teleoperation, in which the human operator directly controls all robot motion in a closed, real-time servo loop. The second strategy is fully autonomous robots.

Teleoperation has improved error-handling functions, as supplied by the human operator. However it has three glaring disadvantages. First, it requires 100% monitoring, since the operator is part of the low-level feedback loops. Second, control is much slower owing to the delays introduced by putting a human in the loop. Third, the robot is vulnerable to inaccurate responses by the human, which are numerous under significant communication latencies.

The autonomy strategy is faster than teleassistance, but suffers by having little error-tolerance. Current real-time robots cannot readily accommodate even simple changes in their environment. If the door handle is different the door-opening experiment fails.

In contrast, teleassistance, which models a layered control structure that uses autonomous routines directed by deictic gestures, is significantly faster than teleoperation, only requires a fraction of the total task time for executive control by the human operator, and can better accommodate natural variations in tasks than can the autonomous routines alone.

4 Example 2: Human cues for object recognition

In teleassistance, economical human gestures bind the robot to a local context. Within that momentary binding, the robot can readily extract and servo on the key features needed to perform the task. All other features in the sensorium can be disregarded or relegated to lower-level controllers (e.g., a damped oscillator). But how does the robot know which features are salient? In the example above, the selection of features is hard-coded in the motor program. This next, proposed experiment looks at how a person could cue the robot to the features that are salient to the current context.

This experiment is work in progress. The intent is to consider linguistic and gestural phenomena in human-computer interaction for the task of visual object recognition. The goal is to discern cues from natural human input that advantageously constrain computation as the computer learns to identify, label, and later recognize and locate objects in a room. The method will exploit spoken phrases by associating syntactic and semantic features with visual features of the object and its surroundings; it will constrain the visual attention to the area pointed to; and it will extract additional delimiting features from iconic gestures. The project addresses issues in linguistics, human-computer interaction, machine learning, computer vision, and gesture understanding.

The platform is the MIT AI Lab Intelligent Room (for details see [Torrance 1995] or our web site at http://ai.mit.edu/projects/hci). The room is an ordinary conference room equipped with multiple cameras, microphones, various displays, a speech synthesizer, and nine dedicated workstations. The room is able to be aware of people, by means of visual tracking and rudimentary gesture recognition, and to be commanded through speech, keyboard, pointing, or mouse.

Background

To recognize an object, computer vision algorithms first analyze features in an image or set of camera views. Which features are salient depends in part on the object. The perception of rhythmic motion, for instance, doesn't help one identify a static object. Saliency also depends on the immediate context. For example, color perception is useful when searching for a particular black chair among a roomful of red ones; it is not helpful when all the chairs are black.

Human supervision

Research in computer vision has been successful in designing algorithms that recognize particular features, such as color or edges. Success is bounded, however, by an inability to dynamically select which features are relevant given the target object and context. One can explore this boundary by including human supervisory input during the recognition process. A person can prioritize the feature search by means of natural speech and gesture. Consider a sample scenario. A person points and asks the room "What is that?" The room agent scans the scene in the indicated direction and looks for a predefined set of features. If the computer is unable to detect an object or selects the wrong one, the person can correct the search strategy by highlighting the salient features verbally, for example by saying "no, it's black".

The interaction establishes constraints on the computational process in several ways. By pointing, the person constrains the search area; the speech syntax delimits an initial feature ordering; and the semantic content highlights key features and provides an error signal on the correctness of the match. Within the momentary constraint system defined by the interaction, the appropriate feature detection algorithms can be applied selectively for the task of object recognition.

We will use this system to extend information query to include a 3D interface. Previously, query systems relied on keyboard input solely. More recently, there

has been work in integrating 2D spatial data in the query, such as by selecting coordinates on a map with a mouse or pointer. In this proposal, one can extend the interface to include queries about physical objects in a room. For example one might point to a VCR and ask what it is and how to operate it.

Sample Object Recognition Scenario

```
1. LEARNING
   Person
      Speech: "That is a VCR"
      Gesture: Pointing
   Computation
      Speech analysis:
        Analyze syntactic structure
        Search for known object "VCR"
        Store discourse context
      Gesture analysis:
        Circumscribe visual attention to the
        direction pointed to
      Visual processing:
        Initialize visual feature map with
        known or discerned parameters
        Visual search within attention cone
   Computer Output
      Speech: Affirm or request clarification
      Gesture: Highlight candidate object if
        found

2. CORRECTION
   Person
      Speech: "No, it's black" (or "bigger",
        "to the left", etc.)
      Gesture: Pointing or Iconic (e.g.,
        describes object shape or size)
   Computation
      Speech analysis:
        Analyze syntactic structure
        Refer to discourse context
      Gesture analysis:
        Adjust attention to new direction
        Recognize iconic gesture and extract
        relevant spatial parameters
      Visual processing:
        Tune visual feature map with new
        parameters
        Repeat visual search
   Computer Output
      Speech: Affirm or request clarification
      Gesture: Highlight candidate object

3. LEARNED
   Person
      Speech: "Right"
   Computation
      Extract other camera views
      Store salient features
      Label room contents database
   Computer Output
      Spoken: affirm

4. RECALL
   Person
      Speech: "What is that?" "Where is the
        VCR?"
      Gesture: Pointing
   Computation
      Search labeled database for feature map
      Visual search
      Select instance nearest to the person
   Computer Output
      Spoken: affirm or request help
      Gesture: Highlight candidate object

5. INFORMATION QUERY
   Person
      Speech: "How does this work?"
      Gesture: Pointing
   Computation
      Match recognized image against database
      template.
   Computer Output
      Visual and verbal instructions on use.
```

5 Conclusion

There is obvious application for human supervision in assistive technology. Less obviously, it allows researchers to consider the integration of existing technologies in more interesting and complex domains than would be permitted if the computer had to operate autonomously. For example, one can study the influence of spoken cues on computer vision without a working model of the cognitive structure that prompts the speech. What visual, oral, and gestural features are important when? What are the essential cues needed by reactive artificial agents? These questions go toward designing communication interfaces for both the disabled and the able-bodied.

References

[1] P. E. Agre and D. Chapman. Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 268–272. Morgan Kaufmann, Los Altos, CA, 1987.

[2] M. Arbib, T. Iberall, and D. Lyons. Coordinated control programs for movements of the hand. Technical report, COINS Department of Computer and Information Science, University of Massachussetts, 1985.

[3] D. H. Ballard, M. M. Hayhoe, P. K. Pook, and R. Rao. Deictic codes for the embodiment of cognition. *The Behavioral and Brain Sciences*, 1996.

[To appear – earlier version available as National Resource Laboratory for the study of Brain and Behavior TR95.1, January 1995, U. of Rochester].

[4] D.H. Ballard, M.M. Hayhoe, F. Li, and S.D. Whitehead. Hand-eye coordination during sequential tasks. *Phil. Transactions of the Royal Society of London*, March 1992.

[5] J. Crisman and M. Cleary. Deictic primitives for general purpose navigation. In *Proceedings of the AIAA Conference on Intelligent Robots in Factory, Field, Space, and Service (CIRFFSS)*, March 1994.

[6] Y. Guiard. Asymmetrical division of labor in human skilled bimanual action: the kinematic chain as a model. *Journal of Motor Behavior*, 19(4):486, 1987.

[7] E. Kowler and S. Anton. Reading twisted text: Implications for the role of saccades. *Vision Research*, pages 27:45–60, 1987.

[8] J. Pelz, M. M. Hayhoe, D. H. Ballard, and A. Forsberg. Separate motor commands for eye and head. *Investigative Ophthalmology and Visual Science, Supplement*, 1993.

[9] P. K. Pook and D. H. Ballard. Deictic teleassistance. In *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems (IROS)*, September 1994.

[10] M.C. Torrance. Advances in human-computer interaction by: The intelligent room. In *Research Symposium, Human Factors in Computing: CHI'95 Conference*, May 1995.