PERCEPTUAL USER INTERFACES

MATTHEW TURK
AND GEORGE ROBERTSON,
Guest Editors

THERE IS NO MOORE'S LAW FOR USER interfaces. Human-computer interaction has not changed fundamentally for nearly two decades. Most users interact with computers by typing, pointing, and clicking. The majority of work in human-computer interfaces (HCI) in recent decades has been aimed at creating graphical user interfaces (GUIs) that give users direct control and predictability. These properties provide the user a clear model of what commands and action are possible and what their affects will be; they allow users to have a sense of accomplishment and responsibility about their interactions with computer applications.

Although these endeavors have been very successful, and the WIMP (windows, icons, menus, pointer) paradigm has served to provide a stable and global face to computing, it is clear this paradigm will not scale to match the myriad form factors and uses of computers in the future. Computing devices are becoming smaller and ubiquitous, and interaction with them is becoming more and more pervasive in our daily lives. At the same time, large-scale displays are becoming more common, and we are beginning to see a convergence between computers and television. In all cases, the need arises for more general and intuitive ways of interacting with the technology. Pointing, clicking, and typing—though still appropriate for many uses of computers in the foreseeable future—will not be how most people interact with the majority of computing devices for long.

What we need are interaction techniques well matched with how people will use computers. From small, mobile devices carried or worn to powerful devices embedded in homes, businesses, and automobiles—one size does not fit all. Is there a paradigm that captures the essence of such diverse future HCI requirements? We believe there is, and it is grounded in how people interact with each other and with the real world. This is the essence of *perceptual user interfaces (PUIs)*.

PUIs are characterized by interaction techniques that combine an understanding of natural human capabilities (particularly communication, motor, cognitive, and perceptual skills) with computer I/O devices and machine perception and reasoning. They seek to make the user interface more natural and compelling by taking advantage of the ways in which people naturally interact with each other and with the world—both verbally and nonverbally. Devices and sensors should be transparent and passive if possible, and machines should perceive relevant human communication channels as well as generate output that is naturally understood. This is expected to require integration at multiple levels of technologies such as speech and sound recognition and generation, computer vision, graphical animation and visualization, language understanding, touch-based

PUI

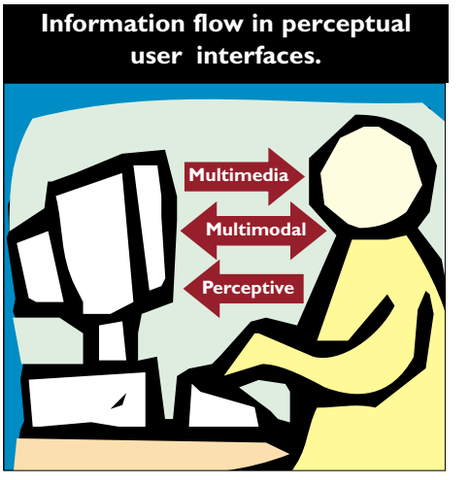
JEAN-FRANCOIS PODEVIN

sensing and feedback (haptics), learning, user modeling, and dialogue management.

The accompanying figure illustrates how PUI encompasses research in several areas. Although the figure shows information flow in the context of a traditional computer form factor, PUI is intended for new form factors as well.

A *perceptive UI* (as opposed to PUI) is one that adds human-like perceptual capabilities to the computer, for example, making the computer aware of what the user is saying or what the user's face, body, and hands are doing. These interfaces provide input to the computer while leveraging human communication and motor skills.

Multimodal UI is closely related, emphasizing human communication skills. We use multiple modalities when we engage in face-to-face communication, leading to more effective communication. Most work on multimodal UI has focused on computer input (for example, using speech together with pen-based gestures). Multimodal output uses different modalities, like visual display, audio, and tactile feedback, to engage human perceptual, cognitive, and communication skills in understanding what is being presented. In multimodal UI, various modalities are sometimes used independently and sometimes simultaneously or tightly coupled.



Information flow in perceptual user interfaces.

Multimedia UI, which has experienced an enormous amount of research during the last two decades, uses perceptual and cognitive skills to interpret information presented to the user. Text, graphics, audio, and video are the typical media used. Multimedia research focuses on the media, while multimodal research focuses on the human perceptual channels. From that point of view, multimedia research is a subset of multimodal output research.

PUI INTEGRATES PERCEPTIVE, MULTIMODAL, AND multimedia interfaces to bring our human capabilities to bear on creating more natural and intuitive interfaces. PUIs will enhance the use of computers as tools or appliances, directly enhancing GUI-based applications—for example, by taking into account gestures, speech, and eye gaze ("No, *that* one"). Perhaps more importantly, these new technologies will enable broad uses of computers as assistants, or agents, that will interact in more human-like ways. Perceptual interfaces will enable multiple styles of interaction—such as speech only, speech and gesture, text and touch, vision, and synthetic sound—each of which may be appropriate in different circumstances, whether that be desktop apps, hands-free mobile use, or embedded household systems.

THERE ARE A NUMBER OF CHALLENGES FACING THE development and use of PUIs. It is an ambitious endeavor with diverse elements. The articles in this special section present both challenges and early results toward the goal of perceptual interfaces. They are not exhaustive, but rather serve as examples of efforts in this area. (See [1] for others.) Oviatt and Cohen summarize multimodal interfaces, emphasizing their extensive work on speech and pen-based systems. This work shows how multiple modalities can lead to more stable and robust systems (for example, reducing error and disfluency rates).

Pentland proposes *perceptual intelligence* as being key to interfacing with the coming generations of machines; he describes smart rooms and smart clothes—two classes of adaptive sensor-based environments—and technologies required to support them. Crowley et al. delve into the specific area of computer vision-based sensing and perception of human activity. They provide a broad view of the field and describe two projects that use visual perception to enhance graphical interfaces. Reeves and Nass address the need to better understand human perception and psychology as it relates to interaction with technology, and describe results from their human-centered experiments. The sidebars by Tan and Picard provide additional information about specific PUI research area, namely haptics and affective computing, while Bobick et al. describe a large-scale PUI application called the "KidsRoom." C

REFERENCES

1. Turk, M., Ed. *Proceedings of the 1998 Workshop on Perceptual User Interfaces;* research.microsoft.com/PUIWorkshop; www.research.microsoft. com/PUI-Workshop.

MATTHEW TURK (mturk@microsoft.com) is a researcher in the Vision Technology Group at Microsoft Research in Redmond, Wash. **GEORGE ROBERTSON** (ggr@microsoft.com) is a senior researcher at Microsoft Research in Redmond, Wash., working on 3D user interfaces and information visualization.

PERCEPTUAL INTELLIGENCE

Alex Pentland

Good-bye keyboard, so long mouse.
Hello smart rooms and clothes that recognize acquaintances, understand speech, and communicate by gesture.
And that's just the beginning. . .

Inanimate things are coming to life. However, these stirrings are not Frankenstein or the humanoid robots dreamed of in artificial intelligence laboratories. This new awakening is more like Walt Disney: the simple objects that surround us are gaining sensors, computational powers, and actuators. Consequently, desks and doors, TVs and telephones, cars and trains, eyeglasses and shoes, and even the shirts on our backs are changing from static, inanimate objects into adaptive, reactive systems that can be more friendly, useful, and efficient. Or, of course, these new systems could be even more difficult to use than current systems; it depends how we design the interface between the world of humans and the world of this new generation of machines.

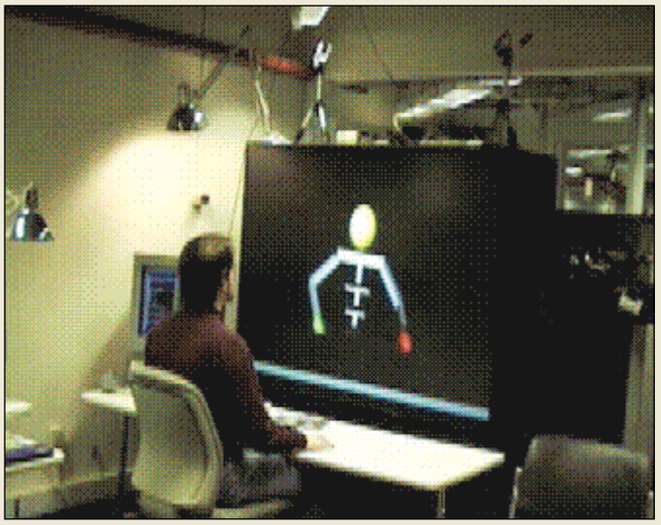
To change inanimate objects like offices, houses, cars, or glasses into smart, active helpmates they need what I call "perceptual intelligence." Translated, perceptual intelligence is paying attention to people and the surrounding situation in the same way another person would, thus allowing these new devices to learn to adapt their behavior to suit us, rather than adapting to them as we do today.

This approach is grounded in the theory that most appropriate, adaptive biological behavior results from perceptual apparatus classifying the situation correctly, which then triggers fairly simple, situation-specific learned responses. It is an ethological view of behavior, and stands in strong contrast to cognitive theories that hold that adaptive behavior is primarily the result of complex reasoning mechanisms.

From this theoretical perspective the problem with current computers is they are incredibly isolated. If you imagine yourself living in a closed, dark, soundproof box with only a telegraph connection to the outside world, you can get some sense of how difficult it is for computers to act intelligently or be helpful. They exist in a world almost completely disconnected from ours, so how can they know what they should do in order to be helpful?

In the language of cognitive science, perceptual intelligence is the ability to deal with the frame problem: It is the ability to classify the

Figure 1. These systems use 2D camera observations to drive a dynamic model of the human's motion. The dynamic model uses a control law that chooses typical behaviors when it is necessary to choose among multiple physically possible trajectories. Predictive feedback from the dynamic model is provided by setting priors for the 2D observation process. These real-time systems have been successfully integrated into applications ranging from Becker's physical rehabilitation trainer (using a 3D model) to Sparacino's computer-enhanced dance space (using 2.5D models) [12]



current situation, so that you know what variables are important, and thus can take appropriate action. Once a computer has the perceptual ability to know who, what, when, where, and why, then I believe that probabilistic rules derived by statistical learning methods are normally sufficient for the computer to determine a good course of action.

The key to perceptual intelligence is making machines aware of their environment, and in particular, sensitive to the people who interact with them. They should know who we are, see our expressions and gestures, and hear the tone and emphasis of our voice. People often confuse perceptual intelligence with ubiquitous computing or artificial intelligence, but in fact they are very different.

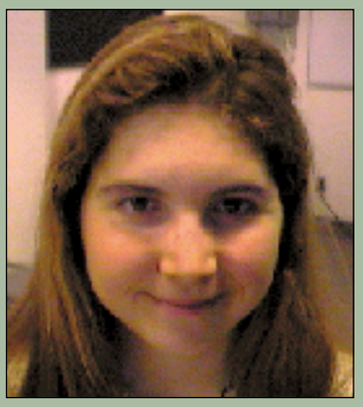
The goal of the perceptual intelligence approach is not to create computers with the logical powers envisioned in most AI research, or to have computers that are ubiquitous and networked, because most of the tasks we want performed do not seem to require complex reasoning or a god's-eye view of the situation. One can imagine, for instance, a well-trained dog controlling most of the functions we envision for future smart environments. So instead of logic or ubiquity, we strive to create systems with reliable perceptual capabilities and the ability to learn simple responses.

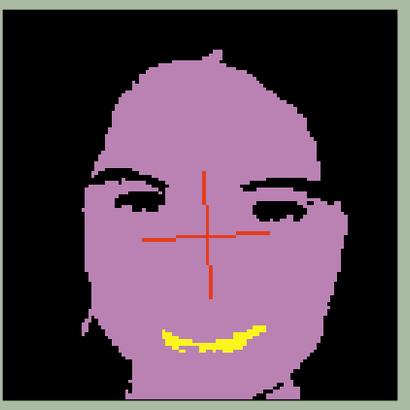
One implication of this approach is we often discover it is not necessary to have a general-purpose computer in the system or to have the system networked together with other resources. In fact, a design goal that my research group usually adopts is to avoid tight networking whenever possible. We feel that ubiquitous networking and its attendant capacity to concentrate information has too close a resemblance to George Orwell's dark vision of a government observing your every move. Instead, we propose that local intelligence—mainly perceptual intelligence combined with relatively sparse, user-initiated networking—can provide the same benefits as ubiquitously networked solutions, while making it more difficult for outsiders to track and analyze user behavior.

A key idea of perceptually intelligent interfaces is they must be adaptive both to the overall situation and to the individual user. As a consequence, much of our research focuses on learning user behaviors, and how user behavior varies as a function of the situation. For instance, we have built systems that learn a user's driving behavior, thus allowing the automobile to anticipate the driver's actions, and a system that learns typical pedestrian behaviors, allowing it to detect unusual events [6].

Most recently, we have built audiovisual systems that learn word meanings from natural audio and visual input [7]. This automatically acquired vocabulary can then be used to understand and generate spoken language. Although simple in its current form, this effort is a first step toward a more fully grounded model of language acquisition. The current system

Figure 2.





(top) The LAFTER system finds and tracks face and facial features at 30Hz, feeding facial feature geometry for expression recognition; (bottom left) accurate, real-time recognition of a 40-word American Sign Language vocabulary; (bottom right) recognizing and teaching T'ai Chi gestures.

can be applied to human-computer interfaces that use spoken input. A significant problem in designing effective spoken word interfaces has always been the difficulty in anticipating a person's word choice and associated intent. Our system addresses this problem by learning the vocabulary choices of each user together with the semantic grounding of the word. This methodology is now used to build several practical systems, including adaptive human-machine interfaces for browsing, education, and entertainment.

To explore this vision of helpful, perceptually intelligent environments my colleagues and I have created a series of experimental testbeds at the MIT. Media Laboratory. These testbeds can be divided into two main types: *smart rooms* and *smart clothes.* The idea of a smart room is a little like having a butler; that is, a passive observer who usually stands quietly in the corner but who is constantly looking for opportunities to help. Smart clothes, on the other hand, act more like personal assistants. They are like a person who travels with you, seeing and hearing everything you do, and trying to anticipate your needs and generally smooth your way.

Both smart rooms and smart clothes are instrumented with sensors that allow the computer to see, hear, and interpret users' actions (currently mainly cameras, microphones, and electromagnetic field sensors, but also biosensors like heartrate and muscle action). People in a smart room can control programs, browse multimedia information, and experience shared virtual environments without keyboards, special sensors, or special goggles. Smart clothes can provide personalized information about the surrounding environment, such as the names of people you meet or directions to your next meeting, and can replace most computer and consumer electronics. The key idea is that because the room or the clothing knows something about what is going on, it can react intelligently.

Our first smart room was developed in 1989; now there are smart rooms in Japan, England, and throughout places in the U.S. They can be linked together by ISDN telephone lines to allow shared virtual environment and cooperative work experiments. Our smart clothes project was started in 1992, and now includes many separate research efforts.

Figure 3. Toco the Toucan. This computer graphics demonstration of word and gesture earning for human-machine interactions was called "one of the best demos at SIGGRAPH '98 " by the *Los Angeles Times*.

Smart Rooms

Here, I describe some of the perceptual capabilities available to our smart rooms, and provide a few illustrations of how these capabilities can be combined into interesting applications. This list of capabilities is far from exhaustive; mostly it is a catalog of our most recent research in each area.[1]

To act intelligently in a day-to-day environment, the first thing you need to know is: *where are the people?* The human body is a complex dynamic system, whose visual features are time varying, noisy signals. Accurately tracking the state of such a system requires use of a recursive estimation framework. The elements of the framework are the observation model relating noisy low-level features to the higher-level skeletal model and vice versa, and the dynamic skeletal model itself.

This extended Kalman filter framework reconciles the 2D tracking process with higher-level 3D models, thus stabilizing the 2D tracking by coupling an articulated dynamic model directly with raw pixel measurements. Some of the demonstrated benefits of this added stability include increase in 3D tracking accuracy, insensitivity to temporary occlusion, and the ability to handle multiple people.

The dynamic skeleton model interpolates those portions of the body state not measured directly, such as the upper body and elbow orientation, by use of the model's intrinsic dynamics and the behavior (control) model. The model also rejects noise that is inconsistent with the dynamic model.

The system runs on a PC at 30Hz, and has performed reliably on hundreds of people in many different physical locations, including exhibitions, conferences, and offices in several research labs. The jitter or noise observed experimentally is 0.9cm for 3-D translation and 0.6 degrees for 3D rotation when operating in a desk-sized environment.

[1]Readers are referred to conferences such as the IEEE International Conference on Automatic Face and Gesture Recognition for related work by other research laboratories.

Figure 4. The author wearing a variety of new devices. The glasses (built by Microoptical, Boston) contain a computer display nearly invisible to others. The jacket has a keyboard literally embroidered into the cloth. The lapel has a context sensor that classifies the user's surroundings. And, of course, there's a computer (not visible in this photograph).

SAM OGDEN

One of the main advantages of feedback from a 3D dynamic model to the low-level vision system. Without feedback, the 2D tracker fails if there is even partial self-occlusion from a single camera's perspective. With feedback, information from the dynamic model can be used to resolve ambiguity during 2D tracking [12].

Once the person is located, and visual and auditory attention has been focused on them, the next question to ask is: *who is it?* The question of identity is central to adaptive behavior because who is giving a command is often as important as the command itself. Perhaps the best way to answer the question is to recognize them by their facial appearance and by their speech.

Face recognition systems in use today are real-time and work well with frontal mug-shot images and constant lighting. For general perceptual interfaces, person recognition systems will need to recognize people under much less constrained conditions.

One method of achieving greater generality is to employ multiple sensory inputs; audio- and video-based recognition systems in particular have the critical advantage of using the same modalities that humans use for recognition. Recent research has demonstrated that audio- and video-based person identification systems can achieve high recognition rates without requiring a specially constrained environment [1].

Facial expression is also critical. For instance, a car should know if the driver is sleepy, and a teaching program should know if the student looks bored. So, just as we can recognize a person once we have accurately located their face, we can also analyze the person's facial motion to determine their expression. The lips are of particular importance in interpreting facial expression, and so we have focused our attention on tracking and classification of lip shape.

The first step of processing is to detect and characterize the shape of the lip region. For this task we developed the LAFTER system [5]. This system uses an online learning algorithm to make maximum a posteriori (MAP) estimates of 2D head pose and lip shape, runs at 30Hz on a PC, and has been used successfully on hundreds of users in many different locations and laboratories. Using lip shape features derived from LAFTER we can train hidden Markov models (HMMs) for various mouth configurations. HMMs are a well-developed statistical modeling technique for modeling time-series data, and are used widely in speech recognition. Recognition accuracy for eight different users making over 2,000 expressions averaged 96.5%

We have used the recovered body geometry for several different gesture recognition tasks, including a real-time American Sign Language reader and a system that recognizes T'ai Chi gestures, and trains the user to perform them correctly. Typically these systems have a gesture vocabularies of 25 to 50 gestures, and recognition accuracies above 95% [9].

In our first systems we used HMMs to recognize hand and body gestures. We found that although HMMs could be used to obtain high accuracy gesture recognition, they also required a labor-intensive period of training. This is because using HMMs to describe multipart signals (such as two-handed gestures) requires large amounts of training data.

To improve this situation, we developed a new

method of training a more general class of HMM, called the "Coupled Hidden Markov Model." Coupled HMM's allow each hand to be described by a separate state model, and the interactions between them to be modeled explicitly and economically. The consequence is that much less training data is required, and the HMM parameter estimation process is much better conditioned [6].

Almost every room has a chair, and body posture information is important for assessing user alertness and comfort. Therefore, our smart chair senses the pressure distribution patterns in the chair and classifies the seating postures of its user (See Tan below). Two Tekscan sensor sheets (each consisting of a 42-

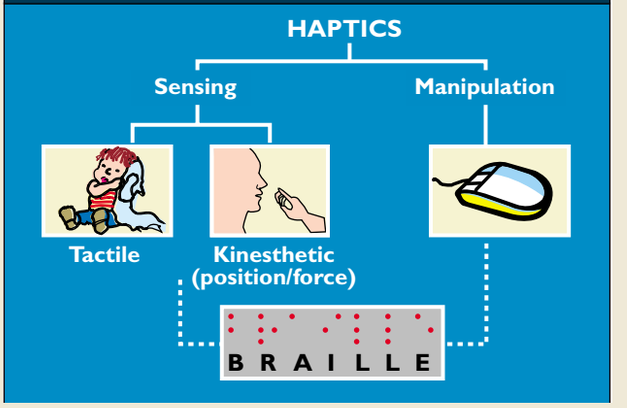
Haptic Interfaces

Hong Z. Tan

Creating interfaces that envelop a sense of touch has met with measured success.

The term "haptics" refers to sensing and manipulation through the sense of touch. Although the word haptics may be new to you, chances are that you're already using haptic interfaces (for example, your keyboard and mouse). As Figure 1 shows, the haptic sensory system (or taction) is usually regarded as having two components: *tactile* (or cutaneous) sensing, and *kinesthetic* sensing (or proprioception). Tactile sensing refers to an awareness of stimulation to the outer surface of the body (the softness of a blanket). Kinesthetic sensing refers to an awareness of limb position and movement (for example, an ability to touch your nose with your eyes closed), as well as muscle tension (for example, estimation of object weights) [1]. Unlike vision and audition that are mainly input systems for the human observer, the haptic system is bidirectional. Many activities, such as the reading of Braille text by the blind, require the use of both the sensing and manipulation aspects of the haptic system.

Figure 1. Definition of haptics.



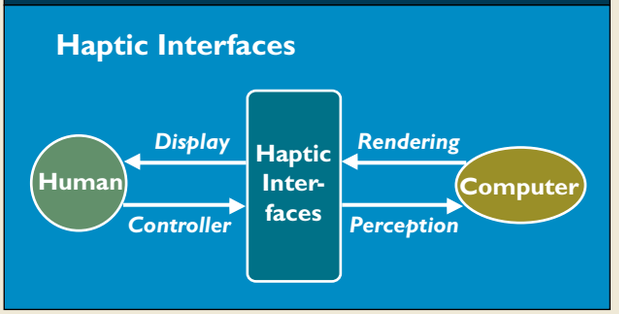
Of the five major human senses—vision, audition, taction, olfaction, and gustation—only the first three have been engaged in most human-machine interface research. Of these three, a disproportional majority of work has been conducted on visual and auditory systems. Historically, work on haptic display has been motivated by the desire to develop sensory-substitution systems for the visually or hearing impaired. Examples include the Optacon (Telesensory Corp., Mountain View, Calif.,), a reading aid for the blind [4]; and TactaidVII (Audiological Engineering Corp., Somerville, Mass.,), a hearing aid for the deaf [6]. These systems can be characterized by an array of vibrators that transform optical or acoustic energy into spatial vibrational patterns. In the past two decades, force-feedback devices (a type of kinesthetic display) have played an important role in teleoperation and virtual reality systems by improving an operator's task performance and by enhancing a user's sense of telepresence. Examples include the Impulse Engine™ (Immersion Corp., San Jose, Calif.) and the popular PHANToM™ (SensAble Technologies Inc., Cambridge, Mass.) [5].

Depending on the direction of information flow (see Figure 2), a human observer would either regard a haptic interface as a display (for example, Optacon, TactaidVII, Impulse Engine, and PHANToM) or a controller (computer mouse). A computer would either render a haptic world through devices such as the PHANToM, or perceive haptic information through contact sensors. An example of a haptic perceptive UI is the sensing chair. Originally conceived at the MIT Media Lab and currently being developed at Purdue University, the sensing chair project is aimed toward a real-time system that tracks the sitting postures of a user through the use of surface-mounted contact sensors (enclosed in the green protective pouches as shown in Figure 3). The realization of a robust tracking system will lead to many exciting applications

by-48 array of force-sensitive resistor units) are mounted to the seatpan and the backrest of the chair and output 8-bit pressure distribution data. This data is collected and the posture is classified using image modeling and classification algorithms.

The current version of the real-time seating posture classification system uses a statistical classification method originally developed for face recognition. For each new pressure distribution map to be classified, a "distance-from-feature-space" error measure is calculated for each of the M postures and compared to a threshold. The posture class that corresponds to the smallest error is used to label the current pressure map, except when all error values exceed the threshold

Figure 2. Information flow with haptic interfaces.



such as automatic control of airbag deployment forces, ergonomics of furniture design, and biometric authentication for computer security.[1]

Despite the progress made in the past two decades (see Srinivasan in [3]), haptic interfaces have not yet become commonplace. One reason, I think, is the technological challenge associated with the design and fabrication of interfaces that make physical contact with human users. This, however, will change as haptic technology matures. The other reason is the lack of killer apps for haptic user interfaces.

Figure 3. The sensing chair is a haptic PUI

To really appreciate the human haptic sensory system requires an understanding of what happens if we are deprived of it. Imagine what happens if one loses the tactile sense. We have all experienced the lack of dexterity with a gloved hand. What happens if one loses the kinesthetic sense? Such cases are rare, but one is well documented in Cole's book on Ian Waterman who, at the age of 19, lost all sensation below his neck [2]. Without that sixth sense of joint and limb positions in space, he fell on the floor in a heap, unable to stand or walk. With sheer courage and determination, Waterman eventually taught himself to walk again by constant visual monitoring of his body position. However, as Cole pointed out, Waterman's new way of walking was like "a wooden puppet activated by a novice." It lacked the grace observed in our movements of walking, dancing, and running.

The fact that Waterman could walk at all with visual feedback alone attests to our ability to accomplish almost any task with vision. The fact that he could no longer walk gracefully suggests to me that perhaps the killer app of haptic interfaces is to make human–computer interactions more intuitive, natural, and above all, graceful. C

[1] It is much easier to change one's appearance or voice than to fake the distance between the ischial tuberosities (sitting bones), something that is readily detectable by our chair sensors.

References

1. Boff, K.R., Kaufman, L., and Thomas, J.P. (Eds.). *Handbook of Perception and Human Performance: Sensory Processes and Perception. Vols. 1 and 2*. Wiley, New York, N.Y., 1986.
2. Cole, J. *Pride and a Daily Marathon*. MIT Press, Cambridge, MA, 1995.
3. Durlach N. I. and Mavor A. S.. *Virtual Reality: Scientific and Technological Challenges*. National Academy Press, Washington, D.C, 1994.
4. Linvill, J.G. and Bliss, J.C. A direct translation reading aid for the blind. In *Proceedings of the Institute of Electrical and Electronics Engineers 54* (1966), 40–51.
5. Massie, T.H., and Salisbury, J.K. The PHANToM haptic interface: A device for probing virtual objects. In *Proceedings of the ASME Dynamic Systems and Control Division*. (1994, Chicago,) American Society of Mechanical Engineers, New York, N.Y., 295–299.
6. Reed, C.M., and Delhorne, L.A. Current results of field study of adult users of tactile aids. *Seminars in Hearing 16*, (1995), 305–315.

Hong Z. Tan (hongtan@ecn.purdue.edu) is an assistant professor in the School of Electrical and Computer Engineering at Purdue University in West Lafayette, Indiana.

Figure 5.

(Top left) The Stochasticks wearable billiards advisor (photo by Sam Ogden); (right) what the Stochasticks user sees; (bottom) a hat-mounted camera observers the user's hands and translates a limited set of American Sign Language into English.

in which case the current posture is declared unknown. The algorithm runs in real-time on a Pentium PC, with a classification accuracy of approximately 95% for 21 different postures.

Traditional interfaces have hard-wired assumptions about how a person will communicate. In a typical speech recognition application the system has some preset vocabulary and (possibly statistical) grammar. For proper operation the user must restrict what is said to words and vocabulary built into the system. However, studies have shown that in practice it is difficult to predict how different users will use available input modalities to express their intents. For example, Furnas et al. did a series of experiments to see how people would assign keywords for operations in a mock interface [2]. They conclude that: "There is no one good access term for most objects...The idea of an "obvious," "self-evident," or "natural" term is a myth! ... Even the best possible name is not very useful...Any keyword system capable of providing a high hit rate for unfamiliar users must let them use words of their own choice for objects." Our conclusion is to make effective interfaces there need to be adaptive mechanisms that learn how individuals use modalities to communicate.

Therefore, we have built a trainable interface, which lets users teach it which words and gestures they want to use and what the words and gestures mean. Our current work focuses on a system that learns words from natural interactions; users teach the system words by simply pointing to objects and naming them.

This work demonstrates an interface that learns words and their domain-limited semantics through natural multimodal interactions with people. The interface, embodied as an animated character named Toco the Toucan, can learn acoustic words and their meanings by continuously updating association weight vectors that estimate the mutual information between acoustic words and attribute vectors representing perceptually salient aspects of virtual objects in Toco's world. Toco is able to learn semantic associations (between words and attribute vectors) using gestural input from the user. Gesture input enables the user to naturally specify which object to attend to during word learning [7]

Smart Clothes

In the smart room, cameras and microphones are watching people from a third-person perspective. However, when we build the computers, cameras, microphones and other sensors into our clothes, the computer's view moves from a passive third person to an active first-person vantage point.

This means smart clothes can be more intimately and actively involved in the user's activities. If these wearable devices have sufficient understanding of the user's situation—-that is, enough perceptual intelligence—then they should be able to act as an intelligent personal agent, proactively providing the wearer with information relevant to the current situation.

For instance, if you build a global position sensor (GPS) into your belt, then navigation software can help you find your way around by whispering directions in your ear or showing a map on a display built into your glasses. Similarly, body-worn accelerometers and tilt sensors can distinguish walking from standing from sitting, and biosensors such as galvanic skin response (GSR) are correlated with mental arousal, allowing construction of wearable medical monitors. A simple but important application for a medical wearable is to give people feedback about their alertness and stress level. More advanced applications, being developed in conjunction with the Center for Future Health at the University of Rochester, include early warning systems for people with high-risk medical problems, and eldercare wearables to help keep seniors out of nursing homes.

These wearable devices are examples of *personalized perceptual intelligence,* allowing proactive fetching and filtering of information for immediate use by the wearer. The promise of such wearable devices recently motivated the IEEE Computer Society to create a Technical Committee on Wearable Information Devices (see iswc.gatech.edu).

While specialized sensors such as GPS, accelerometers, and biosensors may predominate in initial wearable applications, audio and video sensors will soon play a central role. For instance, we have built wearables that continuously analyze background sound to detect human speech. Using this information, the wearable is able to know when you and another person are talking, so that they won't interrupt (imagine having polite cell phones!) Researchers in my laboratory are now going a step further, using microphones built into a jacket to allow word-spotting software to analyze your conversation and remind you of relevant facts.

Cameras make attractive candidates for a wearable, perceptually intelligent interface, because a sense of environmental context may be obtained by pointing the camera in the direction of the user's gaze. For instance by building a camera into your eyeglasses, face recognition software can help you remember the name of the person you are looking at [4, 10].

A more mathematically sophisticated example is to have the wearable computer assist the user by suggesting possible shots in a game of billiards. Figure 5, for instance, illustrates an *augmented reality* system that

helps the user play billiards. A camera mounted on the user's head tracks the table and balls, estimates the 3D configuration of table, balls, and user, and then creates a graphics overlay (using a see-though head-mounted display) showing the user their best shot [3].

In controlled environments, cameras can also be used for object identification. For instance, if objects of interest have bar code tags on a visible surface, then a wearable camera system can recognize the bar code tags in the environment and provide the user with information about the tagged objects [8].

If the user also has a head-mounted display, then augmented reality applications are possible. For instance, by locating the corners of a 2D tag the relative position and orientation of the user and tag can be estimated, and graphics generated that appear to be fixed to the tagged object in the 3D world. Multiple tags can be used in the same environment, and users can add their own annotations to the tag database. In this way, the hypertext environment of the Web is brought to physical reality. Such a system may be used to assist in the repair of annotated machines such as photocopiers or provide context-sensitive information for museum exhibits. Current work addresses the recognition and tracking of untagged objects in the office and outside environments to allow easy, socially motivated annotation of everyday things.

Perhaps just as important but less obvious are the advantages of a self-observing camera. In Figure 5, a downward-pointing camera mounted in a baseball cap allows observation of the user's hands and feet. This view permits the wearable computer to follow the user's hand gestures and body motion in natural, everyday contexts. If the camera is used to track the user's hand, then the camera can act as a direct-manipulation interface for the computer [4, 10]. Hand tracking can also be used for recognizing American Sign Language or other gestural languages. Our most recent implementation recognizes sentence-level American Sign Language in real time with over 97% word accuracy on a 40-word vocabulary [9]. Interestingly, the wearable sign-language recognizer is more accurate than the desk-mounted version, even though the algorithms are nearly identical.

Conclusion

It is now possible to track people's motion, identify them by voice and facial appearance, and recognize their actions in real time using only modest computational resources. By using this perceptual information we have been able to build smart rooms and smart clothes that can recognize people, understand their speech, allow them to control information displays without mouse or keyboard, communicate by facial and hand gesture, and interact in a more personalized, adaptive manner.

We are now beginning to apply such perceptual intelligence to a much wider variety of situations. For instance, we are now working on prototypes of displays that know if you are watching them, credit cards that recognize their owners, chairs that adjust to keep you awake and comfortable, and shoes that know where they are. We imagine building a world where the distinction between inanimate and animate objects begins to blur, and the objects that surround us become more like helpful assistants or playful pets than insensible tools. C

Portions of this article have appeared in *Scientific American* and S*cientific American Presents* and in the *ACM International Symposium on Handheld and Ubiquitous Computing*, 1999.

References

1. Choudhury, T., Clarkson, B., Jebara, T., and Pentland, A. Multimodal person recognition using unconstrained audio and video. In *Proceedings of the Second Int'l Conference on Audio- and Video-based Biometric Person Authentication* (Mar. 22–14, 1999, Washington, DC.), 176–181.
2. Furnas, G., Landaure, T., Gomez, L., and Dumais, S. The vocabulary problem in human-system communications. *Commun. ACM 30,* (1987); 964–972
3. Jebara, T., Eyster, C., Weaver, J., Starner, T., and Pentland, A. Stochasticks: Augmenting the billiards Experience with probabilistic vision and wearable computers. In *IEEE Intl. Symposium on Wearable Computers* (Oct. 23–24, 1997, Cambridge, Mass.).
4. Mann, S. Smart clothing: The wearable computer and WearCam. *Personal Technologies 1,* 1 (1997).
5. Oliver, N., Bernard, F., Coutaz, J., and Pentland, A. LAFTER: Lips and face tracker. *IEEE CVPR '97* (June 17–19, 1997, San Juan, PR). IEEE Press, New York, N.Y.
6. Brand, M., Oliver, N., and Pentland, A. Coupled hidden Markov models for complex action recognition. IEEE CVPR 97. (June 17–19, 1997, San Juan, PR), 994–999. IEEE Press, New York, N.Y.
7. Roy, D., and Pentland, A. Learning words from audio-visual input. In *Proceedings from Int'l Conf. On Speech and Language.* (Dec. 1998, Sydney, Austrialia); 1279.
8. Rekimoto, J., Ayatsuka, Y., and Hayashi, K. Augment-able reality-situated communication through physical and digital spaces. *IEEE Int'l Symposium on Wearable Computers.* (Oct. 19–20, 1998, Pittsburgh); 18–24.
9. Starner, T., Weaver, J., and Pentland, A. Real-time American Sign Language recognition from video using hidden Markov models. *IEEE Trans. Pattern Analy. and Machine Vision.* (Dec. 1998).
10. Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R., and Pentland, A. *Visual Augmented Reality Through Wearable Computing, Presence, Teleoperators and Virtual Environments.* MIT. Press (1997); 163–172.
11. Tan, H., Lu, I., and Pentland, A. The chair as a novel haptic user interface. In *Proceedings of the Workshop on Perceptual User Interfaces (PUI'97).* M. Turk, Ed. (Oct. 19–21, 1997, Banff, Alberta, Canada); 56–57.
12. Wren, C., and Pentland, A. Dynamic modeling of human motion. In *Proceedings for IEEE Face and Gesture Conference* (Nara, Japan, 1998). Also, MIT. Media Laboratory Perceptual Computing Technical Report No. 415.

Alex (Sandy) Pentland (sandy @media.mit.edu) is the Academic Head of the MIT. Media Laboratory, and co-Director of the Center for Future Health. Newsweek magazine recently named him one of the 100 Americans most likely to shape the next century; www.media.mit.edu/~pentland.

All papers and technical reports listed here are available at www.media.edu/vismod.