Perceptive Media: Machine Perception and Human Computer Interaction

Matthew Turk
Computer Science Department
University of California
Santa Barbara, CA 93106 USA
mturk@cs.ucsb.edu

Abstract

Computer hardware has always changed rapidly, but input/output devices, interaction techniques, and software for human-computer interaction have not experienced similar growth and improvement. The GUI-based style of interaction has made computers simpler and easier to use, especially for office productivity applications where computers are used as tools to accomplish specific tasks. However, as the way we use computers changes and computing becomes more pervasive and ubiquitous, largely due to advances in bandwidth and mobility, GUIs will not easily support the range of interactions necessary to meet users' needs. In order to accommodate a wider range of scenarios, tasks, users, and preferences, we need to move toward interfaces that are natural, intuitive, adaptive, and unobtrusive. "Perceptive media" is an interdisciplinary initiative to combine multimedia display and machine perception to create useful, adaptive, responsive interfaces between people and technology. This article describes and investigates aspects of perceptive media and gives examples of work in one particular sub-area, Vision Based Interfaces.

1. Introduction

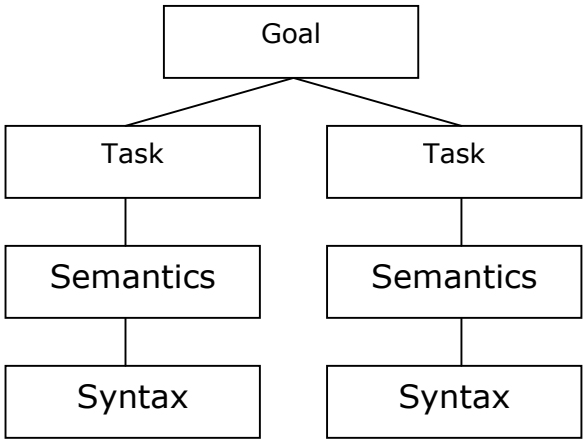
The interface between people and computers has progressed over the years from the early days of switches and LEDs to punched cards, interactive command-line interfaces, and the direct manipulation style of graphical user interfaces. The "desktop metaphor" of graphical user interfaces, a.k.a. WIMP interfaces (for Windows, Icons, Menus, and Pointing devices), has been the standard interface between people and computers for many years. Of course, software and technology for human-computer interaction (HCI) is not isolated from other aspects of computing. Computers have changed enormously over their short history, increasing their speed and capacity, and decreasing component size, at an astounding rate. The size of computers is shrinking, and there are now a plethora of computer devices of various sizes and functionality. In addition, there now are many non-GUI (or "post-WIMP") technologies, such as virtual reality, speech recognition, computer vision, haptics, and spatial sound, that promise to change the status quo in computer-human interaction. But, in general, hardware has changed much more dramatically than software, especially software for HCI.

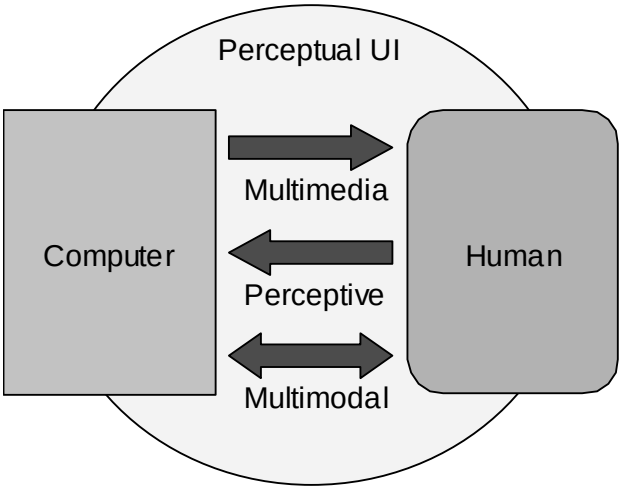
One can view human-computer interaction as a hierarchy of goals, tasks, semantics, and syntax, as shown in Figure 1. The goal level describes what a person wants to do, independent the technology – talk with a friend, for example. Tasks are the particular actions that are required to attain the goal – e.g., locate a telephone, dial a number, talk into the headset. The semantics level maps the tasks onto achievable interactions with the technology, while the syntax level specifies the particular actions (such as double clicking an icon) that accomplish a subtask.

One may view user interfaces are a necessary evil, because they imply a separation between what one wants the computer to do and the act of doing it[1], i.e., a separation between the goal level and the task, semantics and syntax levels. This separation imposes a cognitive load upon the user that is in direct proportion to the difficulty and awkwardness that the user experiences. Poor design, to be sure, exacerbates the problem, giving rise to the all-too-common experience of frustration when using computers.

This frustrating user experience can clearly be improved upon in many ways, and there are many ideas, initiatives, and techniques intended to help, such as user-centered design, 3D user interfaces, conversational interfaces, intelligent agents, virtual environments, and so on.

One point of view is that direct manipulation interfaces – such as the GUI/WIMP model, where users manipulate visual representations of objects and actions – and "information appliances,"[2] devices built to do one particular task well, will alleviate many of the problems and limitations of current computer interfaces. Although this is very likely true – and such devices may well be commercial successes – it is not clear that this interface





style will scale with the changing landscape of form factors and uses of computers in the future.

To complicate things, it is no longer obvious just what "the computer" is; the largely stand-alone desktop PC is no longer the singly dominant device. Rapid changes in form factor, connectivity, and mobility, as well as the continuing effects of Moore's Law, are significantly altering the computing landscape. More and more, computers are embedded in objects and systems that people already know how to interact with (e.g., a telephone or a child's toy) apart from their experience with stand-alone computers.

There are several alternatives for how interacting with computers (whether embedded or not) can proceed in the future, including the following:

1. Simplify – Make the interface obvious and straightforward, giving users direct control and relevant feedback pertaining to the task at hand. Move toward single-task devices and central control to ensure consistency and reliability.

2. Disappear – Make the interface disappear into the device, as with embedded computing (e.g., computer control systems in automobiles), so that users may not even know or care that they are interacting with a computer-based device. A more elaborate version of this is the concept of ubiquitous computing[3,4], where networks of computers, sensors, and displays become intimately integrated into everyday life.

3. Accommodate – Make the interface anticipate, adapt, and react to the user in an intelligent fashion, allowing users to interact in natural ways while the system disambiguates and clarifies users' intentions.

Each of these alternatives has its merits, and each should be (and is being) pursued for future technologies. The first option is the domain of information appliances[2] and direct manipulation interfaces[5,6]. Clearly, the second option is desirable when it is appropriate to the task at hand, as in an automobile braking system – let the embedding computers do their work while the user steps on the brake as he always has done. This seems most useful in traditional uses of computing devices, such as text editing and information query, and in other situations where the computer appears to the user as a *tool* for a specific set of purposes, such as calculating numbers, controlling a process, or drawing.

The third option – interfaces that accommodate to the user in seemingly intelligent or perceptive ways – has developed a significant following in the user interface community in recent years[7,8]. It remains controversial[9], however, and the premise is not yet widely accepted and has not been proven in practice by common commercial systems. For example, anthropomorphism (portraying inanimate computers as having a human-like personality or identity) can be awkward and even confusing to the user[10], although it may also have certain advantages[11]. Speech recognition, the individual technology most associated with this style of interface, has not yet turned the corner to become broadly useful, rather than mildly useful in limited domains. Other component technologies, such as computer vision, reasoning, learning, discourse modeling, and intelligent agents, are still primarily in research labs and have not significantly impacted real systems as of the end of the year 2000. The vision of technology portrayed in the book *2001: A Space Odyssey*[12] is not yet at our disposal.

Nevertheless, one should expect these technologies to mature, especially with the common goal of integrating

them to improve and advance the interface between humans and machines. There is progress every year and hopeful signs that before long they will begin to profoundly affect HCI. In addition to the desire for these technologies to improve the user experience, there is additional motivation for the computer industry: continuing progress in hardware demands more and more software to drive it and consume all those extra cycles.

These three possible directions for HCI development are by no means mutually exclusive; in fact, the second and third have much in common. As people use computers less and less for text-only processing, and more and more for communication and various media-based applications, the future of human-computer interaction becomes completely intertwined with the future of multimedia systems. The two go hand in hand.

2. People and Multimedia

Few individuals have influenced how people use computers more than Vannevar Bush[13] and J. C. R. Licklider[14], with their writings and work toward universal connectivity and interactivity. They articulated visions of how computers can be used to support human activity and creativity in novel ways. Bush's Memex device, for example, shares many characteristics with multimedia and web browsing, but was described four or five decades before these technologies became widely used.

The term multimedia is used broadly to describe just about any combination of media, and more specifically, in the context of personal computers, to describe the simultaneous or alternating use of text, sound, images, and video to present information to users. As we design interfaces to accomplish various tasks in the mobile, heterogeneous environments of the future, all interfaces between people and computers will be multimedia and the available media will expand to include immersive displays, spatial sound, haptic displays, and others. As the presentation of information cannot be considered separately from the receivers of the information, multimedia systems must be studied and developed within the larger context of human-computer interaction, and both directions of information flow (input and output) must be considered and integrated.

In the traditional world of desktop computing, there are "normal users," who comprise the vast majority of the user population, and for whom "one size fits all"; and there are "disabled users," a minority for whom special solutions must be developed. In reality, different people have widely varying needs and preferences, and even individuals have widely varying needs and preferences at different times and in different situations. For example, an interaction style that works well in one situation (e.g., reading text on the desktop monitor) does not fit other situations (e.g., when driving). An individual has different requirements at different stages of life (as a child, a young adult, and an older adult) and in occasional special circumstances (e.g., when a hand or back is injured).

Rather than designing systems that require people to adapt to the technology, we would like to build systems that dynamically adapt to people's needs. This can happen through the use of multimodal systems that understand people – what they do, how they perceive, how they interact – and present information in a flexible manner, rather than a single "take it or leave it" style.

The most natural human interaction techniques are those which we use with other people and with the world around us – that is, those that take advantage of our natural sensing and perception capabilities, along with social skills and conventions that we acquire at an early age. Turk and Robertson[15] described a taxonomy of terms describing such systems, as illustrated in Figure 2:

- *Perceptive User Interfaces* add 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 User Interfaces* are 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 (e.g., 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.
- *Multimedia User Interfaces*, which have had an enormous amount of research during the last two decades, use 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.
- *Perceptual User Interfaces* integrate Perceptive, Multimodal, and Multimedia interfaces to bring our natural human capabilities to bear on creating more natural and intuitive interfaces.

Perceptive media, then, refers to multimedia devices with added perceptual user interface capabilities. These devices integrate human-like perceptual awareness of the environment, especially of the user or users, with the ability to respond appropriately, to adapt to the environment. This requires not only machine perception but also a deep understanding of social conventions such as turn-taking in dialog and non-verbal communication. Progress toward this goal will require research and integration in several areas, including speech/sound recognition, natural language, computer vision, haptics, learning and reasoning, and discourse modeling. It is fundamentally an interdisciplinary endeavor, requiring cooperation between computer scientists and others outside the typical computing fields, such as cognitive science, linguistics, social psychology, and communications. Additionally, common software engineering procedures of design/code/test will not suffice – human-centered design[2] must be embraced, and there will have to be a serious commitment to experimentation and evaluation *in situ*, in real environments. The path is difficult, but the benefits of perceptive media – including natural interaction and liberation from "one size fits all" interaction techniques, such as keyboard and mice – will be significant. This will help to enable universal access to information by all people and in all situations and a more meaningful user experience.

3. Aspects of Perceptive Media

With perceptive media, the information flows in both directions: to the users (e.g., sound and visual displays) and to the computers (e.g., speech and facial expressions of the people nearby). To enable information flow from users to computers in more natural and flexible ways than typing, pointing, and selecting, it is necessary to integrate machine perception technologies into multimedia systems. Of these technologies, two questions arise: (1) which would be most useful, and (2) which are most likely to develop into robust, dependable technologies?

To answer the first question, we can use human-human communication as a model. We grow up interacting with other people daily, almost constantly. The skills and conventions we learn along the way become natural and effortless in most cases. In addition to understanding speech and recognizing people and objects with ease, we rely on social conventions, such as turn-taking in conversations, to guide our actions and reactions. We infer the emotional state of others by perceiving their facial expressions and body language. We communicate directly via gestures, both obvious and subtle (some of which are culturally specific). Perhaps most importantly, we are able to disambiguate information both passively, through understanding the context or background of the conversation, and actively, by querying the other person until the uncertainly is sufficiently reduced.

In order to endow computers with similar capabilities, we need significant progress in several technologies, including:

- Speech and sound recognition
- Natural language understanding
- Computer vision
- Dialog management/planning
- Learning
- User modeling
- Haptics

Although there are vigorous research communities in these areas, as well as in multimedia systems, only in recent years has there begun significant efforts in combining and integrating these technologies in coherent human-computer interfaces. In concert with technological progress, there is a small but growing body of knowledge about how people interact with technology from sociological and psychological points of view – e.g., ways in which people unconsciously attribute human characteristics to computers. According to Reeves and Nass[16], people interact with computers, television, and new media in ways that are *fundamentally social and natural*, just like interactions in real life. For example, people are polite to computers and display emotional reactions to technology.

These findings are not limited to a particular type of media nor to a particular type of person. Such interactions are not conscious – although people can bypass the media equation, it requires effort to do so and it is difficult to sustain. This makes sense, given the fact that, during millennia of human existence anything that appeared to be social was in fact a person. The social responses that evolved in this environment provide a powerful, built-in assumption that can explain social responses to technology – even when people know the responses are inappropriate.

This raises the issue of (although does not explicitly argue for) anthropomorphic interfaces, which are designed to appear intelligent by, for example, introducing a human-like voice or face in the user interface (e.g., in a public kiosk[17]). Schneiderman[5,10,18] argues against anthropomorphic interfaces, emphasizing the importance of direct, comprehensible and predictable interfaces which give users a feeling of accomplishment and responsibility. In this view, adaptive, intelligent, and anthropomorphic interfaces are shallow and deceptive, and they preclude a clear mental model of what is possible and what will happen in response to user actions. Instead, users want a sense of direct control and predictability, with interfaces that support direct manipulation.

Wexelblat[19] questions this point of view and reports on a preliminary study that brings the anti-

anthropomorphic argument into question. The experiment involved users performing tasks presented to them with different interfaces: a "standard" interface and an anthropomorphic interface. In general, the debate on anthropomorphic interfaces has engendered a great deal of (sometimes heated) discussion in recent years among interface designers and researchers. (As Wexelblat writes, "Don't anthropomorphize computers; they hate that!")

This debate may be somewhat of a red herring. When a computer is seen as a *tool* – e.g., a device used to produce a spreadsheet for data analysis – the anti-anthropomorphic argument is convincing. Users would not want a humanoid spreadsheet interface to be unpredictable when entering values or calculating sums, for example, or when moving cells to a different column. However, when computers are viewed as *media* or *collaborators* rather than as tools, anthropomorphic qualities may be quite appropriate. Tools and tasks that are expected to be predictable should be so – but as we move away from office productivity applications to more pervasive use of computers, it may well be that the requirements of predictability and direct manipulation are too limiting.

Concerning the second question, which are the most promising technologies, there are several examples of promising work and prototype systems that may help give indications. For example, the QuickSet system at OGI[20] is an architecture for multimodal integration, and is used successfully for integrating speech and (pen) gesture as users create and control military simulations. Another system for integrating speech and (visual) gesture is described by Poddar, et al.[21], applied to parsing video of a weather report. Another example of tight integration between modalities is in the budding "speechreading" community[22,23]. These systems attempt to use both visual and auditory information to understand human speech – which is also what people do, especially in noisy environments.

4. Vision Based Interaction

Present-day computers are essentially deaf, dumb, and blind. Several people have pointed out that the bathrooms in most airports are smarter than any computer one can buy, since the bathroom "knows" when a person is using the sink or toilet. Computers, on the other hand, tend to ask us questions when we're not there (and wait 16 hours for an answer) and decide to do irrelevant (but CPU-intensive) work when we're frantically working on an overdue document.

Vision is clearly an important element of human-human communication. Although we can communicate without it, people still tend to spend endless hours travelling in order to meet face to face. Why? Because there is a richness of communication that cannot be matched using only voice or text. Body language such as facial expressions, silent nods and other gestures add personality, trust, and important information in human-to-human dialog. We expect it can do the same in human-computer interaction.

Vision based interfaces (VBI) is a subfield of perceptive media which concentrates on developing visual awareness of people. VBI seeks to answer questions such as:

- Is anyone there?
- Where are they?
- Who are they?
- What are the subject's movements?
- What are his facial expressions?
- Are his lips moving?
- What gestures is he making??

These questions can be answered by implementing computer vision algorithms to locate and identify individuals, track human body motions, model the head and face, track facial features, interpret human motion and actions. (For a taxonomy and discussion of movement, action, and activity, see Bobick[24]).

VBI (and, in general, PUIs) can be categorized into two aspects: *control* and *awareness*. Control is explicit communication to the system – e.g., put *that* object *there*. Awareness, picking up information about the subject without an explicit attempt to communicate, gives *context* to an application (or to a PUI). The system may or may not change its behavior based on this information. For example, a system may decide to stop all unnecessary background processes when it sees me enter the room – not because of an explicit command I issues, but because of a change in its context. Current computer interfaces have little or no concept of awareness. While many research efforts emphasize VBI for control, it is likely that VBI for awareness will be more useful in the long run.

The remainder of this section describes VBI projects to quickly track a user's head and use this for both awareness and control (Section 4.1), recognize a set of gestures in order to control virtual instruments (Section 4.2), and track the subject's body using an articulated kinematic model (Section 4.3).

4.1 Fast, Simple Head Tracking

In this section we present a simple but fast technique to track a user sitting at a workstation, locate his head, and use this information for subsequent gesture and pose analysis (see Turk[25] for more details). The technique is appropriate when there is a static background and a single user – a common scenario.

First a representation of the background is acquired, by capturing several frames and calculating the color mean and covariance matrix at every pixel. Then, as live video proceeds, incoming images are compared with the background model and pixels that are significantly different from the background are labeled as

"foreground", as in Figure 3(b). In the next step, a flexible "drape" is lowered from the top of the image until it smoothly rests on the foreground pixels. The "draping" simulates a row of point masses, connected to each neighbor by a spring – gravity pulls the drape down, and foreground pixels collectively push the drape up (see Figure 3(e)). A reasonable amount of noise and holes in the segmented image is acceptable, since the drape is insensitive to isolated noise. After several iterations, the drape rests on the foreground pixels, providing a simple (but fast) outline of the user, as in Figure 3(d).

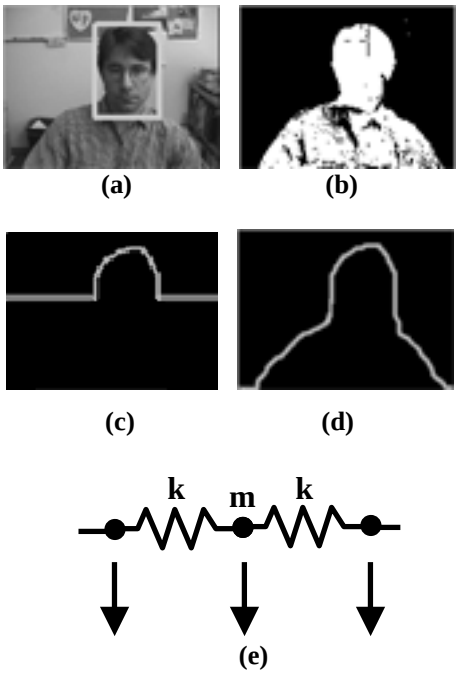


Figure 3. (a) Live video (with head location). (b) Foreground segmentation. (c) Early "draping" iteration. (d) Final "drape". (e) Draping simulates a point mass in each column, connected to its neighbors by springs.

Once the user outline ("drape") settles, it is used to locate the user's head – Figure 3(a) shows the head location superimposed on the live video. All this is done at frame rate in software on a standard, low-end PC. The head location can then be used for further processing. For example, we detect the "yes" and "no" gestures (nodding and shaking the head) by looking for alternating horizontal or vertical patterns of coarse optical flow within the head box. Another use of the head position is to match head subimages with a stored set, taken while looking in different directions. This is used to drive a game of Tic-Tac-Toe, where the head direction controls the positioning of the user's X.

Finally, the shape of the drape (Figure 3(d)) is used to recognize among a small number of poses, based on the outline of the user. Although limited to the user outline, this can be used for several purposes – for example, to recognize that there is a user sitting in front of the machine, or to play a simple visual game such as Simon Says.

4.2 Appearance-Based Gesture Recognition

Recognizing visual gestures may be useful for explicit control at a distance, adding context to a conversation, and monitoring human activity. We have developed a real-time, view-based gesture recognition system, in software only on a standard PC, with the goal of enabling an interactive environment for children[26]. The initial prototype system reacts to the user's gestures by making sounds (e.g., playing virtual bongo drums) and displaying animations (e.g., a bird flapping its wings along with the user).

The algorithm first calculates dense optical flow by minimizing the sum of absolute differences (SAD) to calculate disparity. Assuming the background is relatively static, we can limit the optical flow computation time by only computing the flow for pixels that appear to move. So we first do simple three-frame motion detection, then calculate flow at the locations of significant motion. Once the flow is calculated, it is segmented by a clustering algorithm into 2D elliptical "motion blobs." See Figure 4 for an example of the segmented flow and the calculated flow blobs. Since we are primarily interested in the few dominant motions, these blobs (and their associated statistics) are sufficient for subsequent recognition.

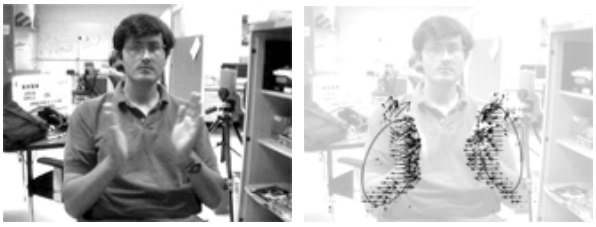


Figure 4. (a) Original image (b) Flow vectors and calculated flow blobs

After calculating the flow blobs, we use a rule-based technique to identify an action. The action rules use the following information about the motion blobs: the number of blobs, the direction and magnitude of motion within the blobs, the relative motion between blobs, the relative size of the blobs, and the relative positions of the blobs. Six actions – waving, clapping, jumping, drumming, flapping, and marching – are currently recognized. Once the motion is recognized, the system

estimates relevant parameters (e.g., the tempo of hand waving) until the action ceases. Figure shows two frames from a sequence of a child playing the "virtual cymbals."

Informal user testing of this system is promising. Participants found it to be fun, intuitive, and compelling. The immediate feedback of the musical sounds and animated characters that respond to recognized gestures is engaging, especially for children. An interesting anecdote is that the child shown in Figure 5, after playing with this system in the lab, went home and immediately tried to do the same thing with his parents' computer.

(a) **(b)**

Figure 5. A user playing the virtual cymbals, with flow blobs overlaid

4.3 Full Body Tracking

To interpret human activity, we need to track and model the body as a 3D articulated structure. We have developed a system[27] which uses disparity maps from a stereo pair of cameras to model and track articulated 3D blobs which represent the major portions of the upper body: torso, lower arms, upper arms, and head. Each blob is modeled as a 3D gaussian distribution, shown schematically in Figure 6. The pixels of the disparity image are classified into their corresponding blobs, and missing data created by self-occlusions is properly filled in. The model statistics are then re-computed, and an extended kalman filter is used in tracking to enforce the articulation constraints of the human body parts.

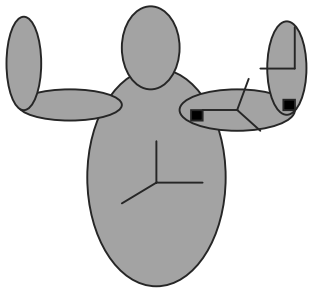


Figure 6. Articulated 3D blob body model

After an initialization step in which the user participates with the system to assign blob models to different body parts, the statistical parameters of the blobs are calculated and tracked. In one set of experiments, we used a simple two-part model consisting of head and torso blobs. Two images from a tracking sequence are shown in Figure 7.

In another set of experiments, we used a four-part articulated structure consisting of the head, torso, lower arm and upper arm, as shown in Figure 8. Detecting and properly handling occlusions is the most difficult challenge for this sort of tracking. The figure shows tracking in the presence of occlusion. Running on a 233 MHz Pentium II system, the unoptimized tracking runs at 10-15 Hz.

Figure 7. Tracking of connected head and torso blobs

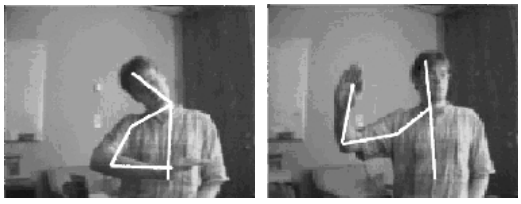


Figure 8. Tracking of head, torso, upper arm, and lower arm

5. Summary

People treat media – including computers, and technology in general – in ways that suggest a social relationship with the media. Perceptive media, modeled after human-to-human interaction, may enable people to interact with technology in ways that are natural, efficient, and easy to learn. A semantic understanding of application and user semantics, which is critical to achieving perceptual interfaces, will enable a single specification of the interface to migrate among a diverse set of users, applications, and environments.

Perceptive media does not necessarily imply anthropomorphic interfaces, although the jury is still out as to the utility of interfaces that take on human-like

characteristics. It is likely that, as computers are seen less as tools for specific tasks and more as part of our communication and information infrastructure, combining perceptual interfaces with anthropomorphic characteristics will become commonplace.

Although the component areas (such as speech, language, and vision) are well researched, the community of researchers devoted to integrating these areas into perceptual media is small – but growing. Some of the critical issues that need to be addressed in the early stages of this pursuit include:

- What are the most relevant and useful perceptual modalities?
- What are the implications for usability testing – how can these systems be sufficiently tested?
- What levels of accuracy, robustness, and integration must machine perceptual capabilities have to be useful in perceptive media?
- What are the compelling tasks ("killer apps") that will demand such interfaces, if any?
- Can (and should) media be introduced in an evolutionary way in order to build on the current GUI infrastructure, or is this fundamentally a break from current systems and applications?

The research agenda for perceptive media must include both (1) development of individual components, such as speech recognition and synthesis, visual recognition and tracking, and user modeling, along with (2) integration of these components. A deeper semantic understanding and representation of human-computer interaction will have to be developed, along with methods to map from the semantic representation to particular devices and environments. In short, there is much work to be done. But the expected benefits are immense.

Acknowledgements

Thanks to Ross Cutler and Nebojsa Jojic for their contributions to this paper. Ross is largely responsible for the system described in Section 4.2. Nebojsa is primarily responsible for the system described in Section 4.3.

References

[1] A. van Dam, "Post-WIMP user interfaces," Communications of the ACM, Vol. 40, No. 2, Pages 63-67, Feb. 1997.

[2] D. A. Norman, *The Invisible Computer*, MIT Press, Cambridge, MA, 1998.

[3] M. Weiser, "The Computer for the Twenty-First Century," Scientific American, September 1991, pp. 94-104.

[4] S. Shafer, J. Krumm, B. Brumitt, B. Meyers, M. Czerwinski, and D. Robbins, "The New EasyLiving Project at Microsoft Research," Proc. Joint DARPA/NIST Smart Spaces Workshop, Gaithersburg, Maryland, July 30-31, 1998.

[5] B. Shneiderman, "Direct Manipulation for Comprehensible, Predictable, and Controllable User Interfaces," Proceedings of IUI97, 1997 International Conference on Intelligent User Interfaces, Orlando, FL, January 6-9, 1997, pp. 33-39.

[6] B. Shneiderman, "The future of interactive systems and the emergence of direct manipulation," *Behaviour and Information Technology*, 1, 1982, pp. 237-256.

[7] M. Maybury, W. Wahlster, *Readings in Intelligent User Interfaces*, Morgan Kaufmann, 1998.

[8] M. Turk (ed.), *Proceedings of the Workshop on Perceptual User Interfaces*, http://www.cs.ucsb.edu/~mturk/PUI/Proc98.

[9] P. Maes, B. Shneiderman, and J. Miller, "Intelligent software agents vs. user-controlled direct manipulation: a debate," *CHI-97 Extended Abstracts: Panels*, ACM, Atlanta, GA, 1997.

[10] B. Shneiderman, "A nonanthropomorphic style guide: overcoming the humpty dumpty syndrome," The Computing Teacher, 16(7), (1989) 5.

[11] A. Wexelblat, "Don't Make That Face: A Report on Anthropomorphizing an Interface," in Intelligent Environments, Coen (ed.), AAAI Technical Report SS-98-02, AAAI Press, 1998.

[12] A. Clark, *2001: A Space Odyssey*, New American Library, 1999 (reissue).

[13] V. Bush, As We May Think, *The Atlantic Monthly*, Volume 176, No. 1, pp.101-108, July 1945.

[14] J. C. R. Licklider and R. W. Taylor, "The computer as a communication device," *Science and Technology*, April 1968.

[15] M. Turk and G. Robertson, "Perceptual User Interfaces," *Communications of the ACM*, March 2000.

[16] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press, September 1996.

[17] K. Waters, J. Rehg, M. Loughlin, S. B. Kang, and D. Terzopoulos, "Visual sensing of humans for active public interfaces," Technical Report CRL 96/5, DEC Cambridge Research Lab, March 1996.

[18] B. Shneiderman, "Beyond intelligent machines: just do it!" *IEEE Software*, vol. 10, 1, Jan 1993, pp. 100-103.

[19] A. Wexelblat, "Don't Make That Face: A Report on Anthropomorphizing an Interface," in *Intelligent Environments*, Coen (ed.), AAAI Technical Report SS-98-02, AAAI Press, 1998.

[20] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow, "QuickSet: Multimodal interaction for distributed applications," *Proceedings of the Fifth Annual International Multimodal Conference*, ACM Press: New York. November, 1997.

[21] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward natural speech/gesture HCI: a case study of weather narration," *Proc. PUI'98 Workshop*, November 1998.

[22] D. Stork and M. Hennecke (eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*, Springer-Verlag, Berlin, 1996.

[23] C. Benoît and R. Campbell (eds.), *Proceedings of the Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, September 1997.

[24] A. Bobick, "Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion," *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, London, England, February 1997.

[25] M. Turk, "Visual interaction with lifelike characters," *Proc. Second IEEE Conference on Face and Gesture Recognition*, Killington, VT, October 1996.

[26] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," *Proc. Third IEEE Conference on Face and Gesture Recognition*, Nara, Japan, April 1998.

[27] N. Jojic, M. Turk, and T. Huang, "Tracking articulated objects in stereo image sequences," submitted 1998.